FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Data Mining for Rational Drug Design

Catarina Isabel Peixoto Candeias

Mestrado em Engenharia Biomédica

Supervisor: Rui Camacho

July 7, 2017

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

MEB - MESTRADO EM ENGENHARIA BIOMÉDICA **2016/2017**

A Dissertação intitulada

"Data Mining for Rational Drug Design"

foi aprovada em provas realizadas em 07-07-2017

o júri

Presidente Prof. Doutor Jorge Alves da Silva
Professor Auxiliar do Departamento de Engenharia Informática da FEUP - U.Porto

Prof. Doutor Carlos Manuel Abreu Gomes Ferreira
Professor Adjunto do Instituto Superior de Engenharia do Porto do Instituto Politécnico do Porto

Prof. Doutor Rui Carlos Camacho de Sousa Ferreira da Silva
Professor Associado do Departamento de Engenharia Informática da FEUP - U.Porto

O autor declara que a presente dissertação (ou relatório de projeto) é da sua exclusiva autoria e foi escrita sem qualquer apoio externo não explicitamente autorizado. Os resultados, ideias, parágrafos, ou outros extratos tomados de ou inspirados em trabalhos de outros autores, e demais referências bibliográficas usadas, são corretamente citados.

Catarina Isabel Peixoto Candeias
Autor - Catarina Isabel Peixoto Candeias

Faculdade de Engenharia da Universidade do Porto

Data Mining for Rational Drug Design

Catarina Isabel Peixoto Candeias

Mestrado em Engenharia Biomédica

Faculdade de Engenharia da Universidade do Porto

July 7, 2017

Resumo

Atualmente existe uma crescente incidência de doenças no mundo e por isso, existe uma maior quantidade de medicamentos disponíveis para promover uma maior eficiência nos sistemas de saúde. Assim, a importância dos novos medicamentos é indiscutível para a vida humana e, consequentemente traduz-se numa maior competência a nível profissional, trazendo muitos benefícios para a sociedade em geral. Por acréscimo, o seu uso consciente conduz também a uma decrescente necessidade de outros cuidados de saúde mais prolongados e de custos mais elevados.

Os fármacos desencadeiam um efeito terapêutico que proporciona uma melhoria de qualidade de vida, no entanto, para sintetizar um fármaco novo, a indústria farmacêutica tem que percorrer um longo, complexo e oneroso processo.

Um dos problemas de saúde que tem vindo a crescer exponencialmente corresponde às doenças neurodegenerativas. Este crescimento veio então aumentar a necessidade de descobrir e desenvolver novos fármacos que possam combater este problema. O processo de conceção de fármacos para este tipo de doenças, em que o fármaco tem que alcançar o Sistema Nervoso Central (SNC), é ainda mais demorado devido à complexidade do cérebro, à tendência dos fármacos para provocarem efeitos adversos graves e principalmente devido à existência da Barreira Hemato-Encefálica (BHE).

O processo de conceção de fármacos é constituído por diversas etapas até atingir a fase final que corresponde aos testes clínicos. Uma das fases do processo testa cinco propriedades importantes de um medicamento, conhecida como fase dos testes ADMET (Absorção, Distribuição, Metabolismo, Excreção e Toxicidade). Estes testes são geralmente muito dispendiosos pois são efetuados em animais.

Com base no historial dos testes ADMET e informação sobre a estrutura e propriedades da molécula do princípio ativo de um fármaco, a informática pode dar um importante contributo para atenuar o problema dos custos elevados e do tempo gasto no processo de desenho de fármacos. Um dos pontos em que o Data Mining pode ser relevante é em evitar ou reduzir a fase em que as moléculas mais promissoras são testadas em animais, ou seja, com base nos resultados anteriores, o Data Mining pode ser utilizado para prever que moléculas vão apresentar resultados mais viáveis.

O trabalho desta dissertação consiste no estudo de técnicas de Data Mining e avaliação das suas potencialidades para melhorar o processo de desenvolvimento de novos fármacos, e sobretudo contribuir para o melhoramento e redução dos custos dos testes ADMET.

Particularmente, consiste na utilização de dois conjuntos de dados: dados com informação sobre a toxicidade das moléculas e dados com informação sobre moléculas que conseguem ultrapassar a Barreira Hemato-Encefálica.

Abstract

Currently there is an increasing incidence of diseases in the world and therefore, there is a greater quantity of medicines available to promote a greater efficiency in the health systems.

Thus, the importance of new medicines is indisputable for human life and, consequently translates into greater professional competence, bringing many benefits to society. In addition, its conscious use also leads to a decreasing need for other, longer and more expensive health care.

Drugs trigger a therapeutic effect that provides an improvement in quality of life. However, to synthesize a new drug, the pharmaceutical industry has to go through a long, complex and costly process.

One of the health problems that has been growing exponentially corresponds to neurodegenerative diseases. This growth increased the need to discover and develop new drugs that could counteract this problem. The drug design process for this type of disease, in which the drug has to achieve the Central Nervous System (CNS), is even more time-consuming due to the complexity of the brain, the tendency of the drugs to cause serious adverse effects and mainly due to the existence of the Blood-Brain Barrier (BBB).

The process of drug design consists of several steps until reaching the final phase that corresponds to the clinical tests. One of the process steps consists on testing five important properties of a drug, known as ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) tests. These tests are usually very expensive because they are done on animals.

Based on the ADMET tests historical and information on the structure and properties of a drug's active principle molecule, Informatics can make an important contribution to alleviating the problem of high costs and time spent in the drug design process. One of the points in which Data Mining may be relevant is to avoid or reduce the phase where the most promising molecules are tested on animals. Based on the previous results, Data Mining can be used to predict which molecules are going to present more viable results.

The work of this dissertation consists of the study of Data Mining techniques and evaluation of its potentialities to improve the process of development of new drugs, and above all to contribute to the improvement and reduction of the costs of ADMET tests. In particular, it consists of the use of two data sets: data with information about the toxicity of molecules and data with information on molecules that can overcome the Blood-Brain Barrier.

Acknowledgment

First of all, I would like to thanks to my supervisor Rui Camacho for all the support that he gave me during this project. Mainly I would like to thank him for all the availability, patient and support shown. I'm sure that without his help and knowledge, it would be more difficult for me.

Secondly, I would like to thanks my family, specially my mother, for giving me the encouragement and strength that I needed to achieve this.

Generalizing, I would like to express my sincere gratitude to those who, directly or indirectly, have contributed to achieving this work.

Finally, I want to thank the project "NanoSTIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics / NORTE-01-0145-FEDER-000016" funded by Northern Regional Operational Program (NORTE 2020), under the agreement of PORTUGAL 2020, and through the European Regional Development Fund (European Regional Development Fund (ERDF) for the provision of data used to the realization of this project.

Catarina Isabel Peixoto Candeias

"Recomeça... se puderes, sem angústia e sem pressa e os passos que deres, nesse caminho duro do futuro, dá-os em liberdade, enquanto não alcances não descanses, de nenhum fruto queiras só metade."

Miguel Torga

Contents

List of Figures

List of Tables

Abbreviations and Symbols

ABI	Adaptive Business Intelligence
ADMET	Absorption, Distribution, Metabolism, Excretion and Toxicology
AI	Artificial Intelligence
AUC	Area Under the Curve
BBB	Blood-Brain Barrier
ChEBI	Chemical Entities of Biological Interest
CML	Chemical Markup Language
CMS	Common Maximum Substructure
CNS	Central Nervous System
CPDBAS	Carcinogenic Potency Database Summary
CRISP-DM	Cross-Industry Standard Process of Data Mining
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DM	Data Mining
DOGMA	Developing Ontology-Grounded Methods and Applications
DSSTox	Distributed Structure-Searchable Toxicity
EPAFHM	EPA Fathead Minnow
FDA	Food and Drug Administration
ILP	Inductive Logic Programming
InChI	International Chemical Identifier
IUPAC	International Union Of Pure and Applied Chemistry
KDD	Knowledge Discovery in Databases
KIF	Knowledge Interchange Format
k-NN	k-Nearest Neighbors
MAE	Mean Absolute Error
MDL	Molecular Design Limited
MED	Mid-continental Ecology Division
NLM	National Library of Medicine
NMR	Nuclear Magnetic Resonance
NTP	National Toxicology Program
OWL	Web Ontology Language
PACT-F	Preclinical And Clinical Trials knowledge base on bioavailability
PKKB	PharmacoKinetics Knowledge Base
RDD	Rational Drug Design
RDF	Resource Description Framework
RMSE	Root Mean Squared Error
SAR	Structure-Activity Relation
SDF	Structure Data Format
SMILES	Simplified Molecular Input Line Entry System

SVM	Support Vector Machine
TEHIP	Toxicology and Environmental Health Information Program
VMD	Visual Molecular Dynamics
WEKA	Waikato Environment for Knowledge Analysis
W3C	World Wide Web Consortium
WHO	World Health Organization
WWW	World Wide Web
XML	eXtensible Markup Language

Chapter 1

Introduction

In this chapter, a context for the work is provided, the domain problem is identified and the main goals of the dissertation are specified.

1.1 Contextualization

In recent years, an enormous number of drugs has contributed to an increase in efficiency in health systems, promoting an improvement in quality of life of the population and also a decrease in the mortality rate.

Drugs are molecules that when introduced into the human body produce a therapeutic effect. However, the impact of drugs use causes some controversy in society. On the one hand, drugs may increase the average life time expectancy of people and allow curing some diseases leading to an improved quality of life. On the other hand, they can increase health-care costs if used incorrectly, and also lead to adverse effects [1].

The discovery of new drugs is related to scientific and technological innovations. The major advances in some areas such as chemistry and biology, and a better understanding of targets and molecular mechanisms that lead to the onset of diseases, have made possible the discovery of remarkable therapeutic formulations [1].

One drawback of the design of new drugs is that it is a process highly complex, long (about five to twelve years) and costly [1]. When the process refers to the design of drugs for the Central Nervous System (CNS), it can take even longer (up to 16 years).

Regarding the success rate of the drugs for CNS being approved, the percentage is much lower, about 8%, while drugs for the cardiovascular system, for example, have a success rate of 20%[1].

Furthermore, the discovery of new drugs is not always achieved. Normally the main cause for that is the lack of efficacy, the existence of toxicity in preclinical trials and any uncertainty in clinical trials [5]. Rational Drug Design (RDD) is an interdisciplinary process that requires the collaboration of researchers with very different skills [6]. RDD aims at a fast and cheap

[1] Alzheimer's Drug Discovery Foundation. Available in www.alzdiscovery.org/, accessed last time in 22-11-2016

development of new drugs and can strongly benefit from two research areas: Chemoinformatics and Knowledge Discovery in Databases (KDD)[2].

With the advancing of times, technologies using computers play a crucial role in all health-related areas. The increase of technological knowledge leads to greater development in this area. Although Informatics tools can be quite useful in more than one step of the drug design process, the focus of the study will be the contribution that Informatics, and Data Mining specifically, can do to the ADMET tests phase.

There are currently several computational methods to predict drug efficacy. These predictive tests help health care providers in choosing the best drug for each situation, and may avoid future problems and adverse effects in patients. With the use of these tools, the time required for the drug development process decreases substantially.

Data Mining uses historical data to construct predictive models. Since the amount of drug design related data in the Internet has increased at a very large pace in the last years, those data repositories can be used to improve KDD.

1.2 Objectives

The main goal of this dissertation is the use of Chemoinformatics and Data Mining tools in order to help in the KDD process which will lead to a reduction in time and cost of the development of new drugs. The approach adopted in this study is directed to a testing phase, in the process of development of new drugs, related to toxicity (ADMET tests) and is also directed to molecules behaviour for Blood-Brain Barrier penetration. Thus, several key points have to be studied and addressed.

In the first part of this project, the main objectives are to understand the basic concepts of drugs, molecular descriptors and their tools, and the overall process of RDD. Secondly, the goal is to explore the concepts of Data Mining tasks and tools, and describe some related work.

After this research, it is essential to investigate and explore Data Mining tools for predicting drug efficacy. The purpose is to use DM techniques in a simplified way to construct models to predict drug efficacy, in order to reduce the time spent designing a new drug.

Thus, it is extremely necessary to address the following two research questions:

H1. Is DM useful to improve the drug design process?

H2. Can we provide extra domain knowledge to improve the DM construct model?

The answer to these questions is debated in the conclusions of the work in Chapter 5.

[2]KDD is a several step data analysis process, one of which is Data Mining (DM), where models are constructed. Since Data Mining is one of the most important steps, KDD is often referred to as DM. From now on it will be used DM to refer to both the model construction step and the KDD process as a whole.

1.3 Dissertation structure

This report is divided into five chapters. This first chapter presents a general approach and contextualization of the theme, as well as the main goals of the work.

Chapter 2 introduces the basic concepts related to the state of art of the domain: drugs, molecular descriptors and Chemoinformatics tools. In that chapter are also explained several concepts concerning the RDD process, Data Mining algorithms, methodologies and tasks, and ontologies.

In Chapter 3 is presented the project implementation from the data preparation to the processes and algorithms used.

Chapter 4 describes the experiences, results achieved and some conclusions of these.

Finally, the conclusions and future work are presented in Chapter 5.

Chapter 2

State of art

This chapter introduces the reader to the state of the art of the domain of our study. It addresses the essential topics related development of new drugs and Chemoinformatics. It presents the overall process of Rational Drug Design (RDD), a description of the state of the art of the Data Mining and related work on the topic of the dissertation. A set of relevant concepts concerning ontologies are also described.

2.1 Development of new drugs

2.1.1 Basic concepts

Since ancient times, people seek in nature resources to relieve and treat the diseases that arise. In early days, therapeutic resources used were derived from plants and animals, among others. Later on, the search for active principles[1] of plants to create the first drugs[2] improved matters[3]. Most drugs that are used in modern medicine, are the result of the progress achieved since the second world war in the field of synthetic organic chemistry and biotechnology[4]. The process of research and drug development has undergone several changes over time with the advances in molecular biology. A drug is defined as any substance other than food or an artefact, which is used for diagnosis, alleviation, treatment and cure of diseases, as well as for the prevention of them[4]. It can be defined as a chemical that interacts with a body part, to change an existing physiological or biochemical process, and may decrease or increase the function of an organ, tissue or cell, but it does not create new functions[4], [1].

[1] An active principle is defined as the constituent of a drug that is widely responsible for conferring pharmacological effect. Different drugs may have the same active principle.

[2] The concept of drug and medicine sometimes is confused. However, these are different concepts, since the medicine is the final product containing the active ingredient (drug), presented in various pharmaceutical forms (capsule, liquid, among others.)

[3] "Conceitos gerais sobre medicamentos".Available in www.anvisa.gov.br/hotsite/genericos/profissionais/conceitos.htm, accessed last time in 25-10-2016

[4] "Generalidades sobre os fármacos".Available in www.manuaismsd.pt/?id=31, accessed last time in 25-10-2016

Drugs classification

There are several criteria to classify drugs as is possible to see in Figure 2.1. They can be classified according to:

Origin

i) Natural source: inorganic, animal, vegetable (more prevalent).

ii) Synthetic source.

iii) Intermediate: corresponding to products of fermentation and genetic engineering.

Action mode

i) Etiological drugs: treat the cause of a disease; almost all belong to the class of chemotherapeutics used to treat infections and parasitic diseases.

ii) Replacement drugs: to overcome for the deficiency of a substance; this deficiency may be due to poor diet, or physiological disorders; substitution treatment may be temporary or permanent.

iii) Symptomatic drugs: alleviate the symptoms of a disease; they are used to attenuate or neutralize disorders resulting from a pathological condition.

Nature of the disease: classification adopted by World Health Organization (WHO) that distinguishes the drugs by the organ in the body which they operate.

Chemical structure: allows for screening drugs analogs derived from the same compound, which facilitates the establishment of correlations between structure and activity[5].

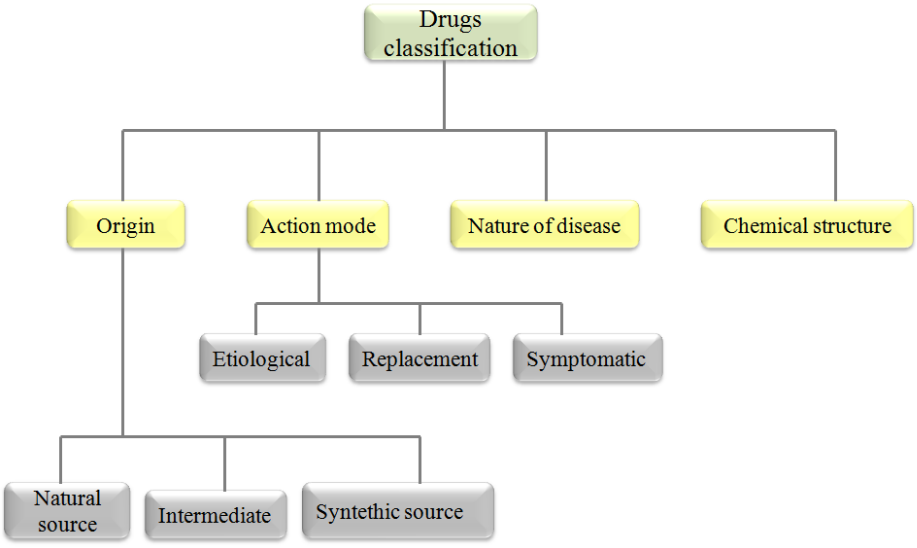


Figure 2.1: Schematic representation of possible criteria to classify drugs. Adapted from[5].

[5] "Classificação de fármacos".Available in https:fenix.tecnico.ulisboa.pt/downloadFile/3779571315641/c, accessed last time in 25-10-2016

2.1.2 Rational Drug Design

Drugs are used since ancient times for catalyzing chemical changes in the human body and for its development. With the widespread use of computers, a new approach for drug design emerged.

Rational Drug Design (RDD) is the process of creating new drugs with the aid of computational methods. This design is based on information about chemical structures for computationally aided analysis and identification of chemical groups that are drug candidates [1].

Over time, drug design has been evolving into more organized processes that become crucial, and new techniques have been tested and used accordingly. After being made a market survey on the disease to be treated, the second step consists in identifying the biological target. The knowledge of the biological target is essential for drug design, because the developed molecules must comply with some minimal structural characteristics to bind to the active regions of the final target. Then it is important to identify the cellular localization of the target at the level of the cell membrane or intracellular. The initial prototype can be found in two distinct ways: based on the binder, that is, when there is prior knowledge of molecules that bind to the desired biological target molecule; and based on the structure, which means that, by methods such as x-ray crystallography or Nuclear Magnetic Resonance (NMR) spectroscopy, the three-dimensional structure of the target is known[6], [7].

The design of the drug structure will depend on the organ it will act, so there are some structural requirements. In addition to bind to the target, the drug must be able to interact with it so, a further step designated by screening, is required.

In the screening phase, different biological assays in molecular, cellular and organic levels, are made to define the activity and selectivity of the future drug candidate.

After these steps, the drug candidates that successfully surpass the screening procedures must be optimized. This optimization can be done by increasing the degree of activity, increasing the selectivity and passing successfully the tests of Absorption, Distribution, Metabolism, Excretion and Toxicology (called ADMET) [1].

One of the major problems with the ingestion of certain drugs is their side effects. Instead of binding to the biological target as intended, drugs can do the same with other undesirable molecules and then trigger chemicals processes which are not important for the treatment of disease and may even raise other problems. So, it is important to reach a prototype with high selectivity, whose activity is reduced to the undesirable molecules [1].

The next step is to conduct ADMET tests in animals to assess each of the characteristics. The last stages of the drug design process are the clinical trials. These are conducted in humans and aim to strengthen the results about the safety of the drug, the recommended dosage and its effectiveness [1].

These clinical studies can be divided into four distinct phases:

[6]Biopharmaceutical Research Development.Available in http://phrma-docs.phrma.org/sites/default/files/pdf/rd_brochure_022307.pdf, accessed last time in 18-10-2016

Phase I: the main objective of this phase is to evaluate the tolerance in humans and the recommended dosage. The patients are monitored for 24 hours, because the purpose is to evaluate the effect of the first dose.

Phase II: this phase aims to study the therapeutic efficacy of the dose range, the kinetics and metabolism.

Phase III: is intended to test the efficacy and safety over a large number of samples, which means, increasing the diversity of people in the tests. The studied drug is administered to a patient sample as similar as possible to the intended population after marketing.

Phase IV: as phase III, this phase aims to test the effectiveness and safety through a high number of samples. After the sale, the drug continues to be studied by pharmacovigilance, which aims to obtain information on their effects, their interactions with other medications and assess their safety. This phase belongs to the responsibility of the regulatory organization and corresponds to the drug study used in the medical practice [1].

Figure 2.2 summarizes all process of development of new drugs.

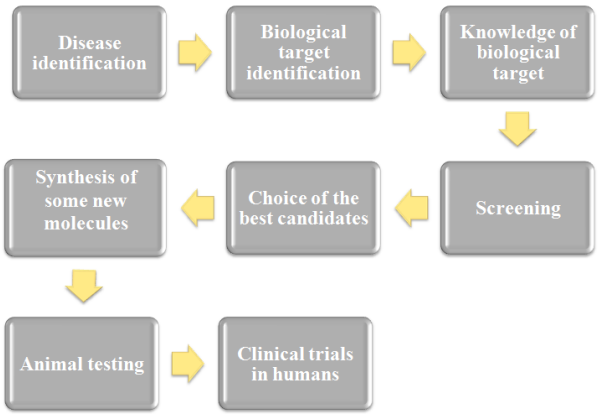


Figure 2.2: Different steps of the development of new drugs. Adapted from [1].

ADMET tests

The human body has many defence mechanisms to expel foreign bodies and to protect it from viruses, infections, etc. The five features of ADMET tests must be understood in order to develop a successful drug [1].

Absorption determines the ability of a compound to overcome the barriers of the human body and reach the target tissue or organ.

Distribution of a drug is usually through the bloodstream, and therefore, the effect of the drug on the target is related to plasma concentration. Different tissues and organs have different absorption capacities and do not have the same rate of blood flow. So, distribution determines the amount of drug which is administered.

Metabolism consists in enzymatic reactions to which the compounds are subject when they enter the organism. These reactions can be hydrolysis- that will inactivate the compound or make it more soluble- or oxidation/reduction reactions that will alter the chemical structure of the compound.

Excretion is the elimination of metabolites (resulting from the breakdown of molecules entering the body) through urine, faeces or sweat.

Finally, the toxicity is the degree of damage that the drug can cause to the organism [1].

Drug design and the Blood-Brain Barrier

Designing drugs to act within the brain is a special case of the design process. The drug design process was already explained, but when referring to Central Nervous System (CNS) drugs, the success rate is much lower. This low rate has several causes, such as, the high degree of complexity of the brain, the problem of the side effects that drugs can cause in the CNS and the presence of the Blood-Brain Barrier (BBB) [8].

Blood-Brain Barrier is a highly selective permeable structure that protects the CNS[7]. This barrier consists of endothelial cells that form capillaries in the brain, and limits the entry of molecules due to the tight junctions formed by transmembrane proteins, negative polarity of the surface and the high level of efflux transporters. The tight junctions significantly reduce the permeation of ions and other small hydrophilic solutes through intercellular gap (paracellular route), thereby forming the physical barrier [9]. This barrier is a vital element in regulating the stability of the internal environment of the brain [10]. Thus, the BBB is very important for the CNS connection to peripheral tissues, and acts as an interface that limits and regulates the exchange of substances between the CNS and the blood. Apart from the selective permeability functions that protect the CNS from damage, BBB has other functions such as providing a stable environment in neural function and maintaining an optimal ionic composition for synaptic signalling function by junction specific ion channels and transporters. The BBB also helps to maintain the separation between the central and peripheral transmitters, thereby reducing their communication [11, 12, 13].

This barrier excludes from the brain all large molecules and 98% of small-molecules of drugs [14]. Only small molecules with high solubility in lipid and a molecular weight of less than 400-500 Daltons can cross the BBB. Unfortunately, only a few brain diseases respond favorably to these drugs. Another problem of small molecules is that only a low percentage of them can cross the BBB in sufficient pharmacological quantities [15].

Drugs that are specific to the CNS must first cross the BBB. The clarification of drug transport mechanisms across the Blood Brain Barrier is important for improving the effectiveness of drugs into the CNS and reducing their toxicity [2, 16, 17].

The transport mechanisms through the BBB can occur in five ways and are represented in Figure 2.3.

[7] "Barreira hematoencefálica".Available in www.oncoguia.org.br/conteudo/o-que-e-a-barreira-hematoencefalica/5720/773/, accessed last time in 25-10-2016

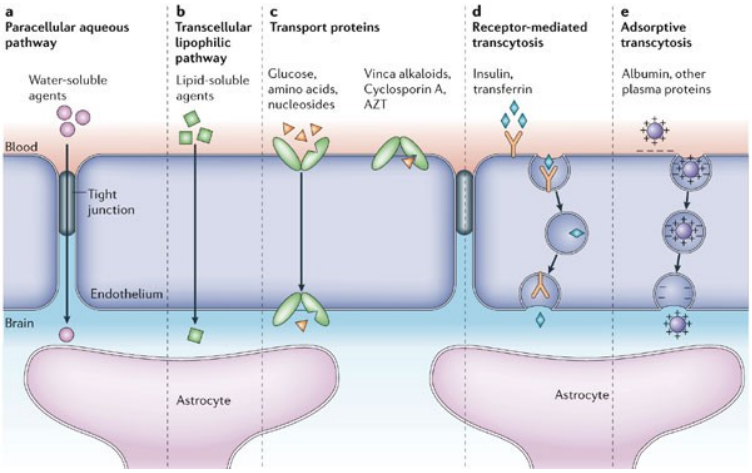


Figure 2.3: Five transport routes through BBB. Extracted from [2].

One of these routes of transport (represented by letter 'a' in the figure), is the **paracellular** and it corresponds to the diffusion of the polar solutes by the tight junctions.

Represented by letter 'b' is **transcelular lipophilic**. This route represents the passive diffusion of lipid soluble agents and is related to some properties of the molecules.

The third route corresponds to the **transport proteins** (represented by letter 'c'). These proteins are in the endothelium and correspond to the entry of various nutrients and endogenous compounds.

Another pathway is **receptor-mediated transcytosis** (represented by letter 'd' in the figure). It corresponds to the transport to the brain of some proteins, such as, insulin and transferrin.

Represented by letter 'e', the fifth route is **adsorptive transcytosis**. There are native plasma proteins that are surrounded by positive charges, in order to increase the transport through this pathway.

BBB is very protective in nature. Thus, the inability of molecules of a drug to penetrate it is a significant impediment for the CNS drug candidates and must be taken into consideration early in the drug design process. This can be changed by an effort to develop knowledge concerning the BBB transport properties and molecular and cellular biology of the capillary endothelium of the brain. However, even if some pharmaceutical company decides to develop a BBB program there are few scientists with knowledge on the subject [2, 9, 14, 16].

2.2 Molecular descriptors

The properties of a molecule contain all of the essential chemical information about it. However, only part of this information is extracted from experimental measurements, as the properties of a molecule do not result from the sum of the properties of its components. Due to this complexity, molecular structures cannot be represented by a single formal model. Thus, various molecular

representations exist that can represent the same molecule[8], [9]. Therefore, to facilitate this representation molecular descriptors were introduced. Molecular descriptors are designed to assist in the drug study process[10].

V. Consonni and R. Todeschini[8] describe a molecular descriptor as "the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment."

Thus, a molecular descriptor is the numerical description of some property of the molecule. These numbers represent various parts of the chemical information contained in different molecular representations. These descriptors are usually used to establish quantitative relationships between structure and properties, and biological activities[11].

Currently, molecular descriptors play an essential role in scientific research, as can be seen in Figure 2.4.

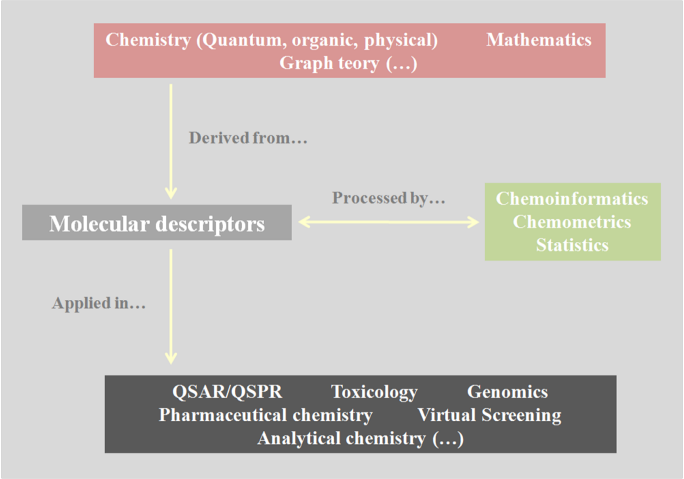


Figure 2.4: Representation of molecular descriptors origin, processing and applications. Adapted from [10].

A proof of the interest of the scientific community by molecular descriptors is the number of descriptors that exist today. About 2000 descriptors are defined and can be calculated using specific software. The different descriptors are distinguished by the complexity of the encoding information and the time required for calculation. Generally, computational requirements increase with what degree of discrimination is achieved, for example, the molecular weight does not transmit

[8] What is a molecular descriptor, Viviana Consonni and Roberto Todeschini, 2000.Available in www.moleculardescriptors.eu/tutorials/T1_moleculardescriptors_what_is.pdf, accessed last time in 25-10-2016

[9] Molecular descriptors: An introduction, 2006. Available in http://infochim.u-strasbg.fr/CS3/program/material/Todeschini.pdf, accessed last time in 24-11-2016

[10] Molecular descriptors and chemometrics: a powerful combined tool for pharmaceutical, toxicological and environmental problems, Roberto Todeschini.Available in www.iamc-online.org/tutorials/T2_moleculardescriptors_chemom.pdf, accessed last time in 25-10-2016

[11] An integrated web-based platform for molecular descriptor and fingerprint computation, March, 2015. Available in www.scbdd.com/chemdes/, accessed last time in 17-10-2016

much on the properties of a molecule but is very quick to compute, while the quantum mechanics can provide accurate representations of properties but take longer time to be calculated [18, 19].

There are several types of molecular descriptors:

3D descriptors: depend on the internal coordinates or absolute orientation of the molecule. 3D descriptors encode several aspects of the three-dimensional structure of the molecule. These descriptors are used in many studies because of the relevance that accompanies the relationship between the ligand conformation and its bioactivity;

2D descriptors: represent descriptors of single value calculated from the graph of the molecule and characterize the structures according to the size, degree of branching and the shape of the overall molecule;

0D and 1D descriptors: are very general, fast to calculate, and therefore do not describe sufficiently the molecule;

fingerprints: are a particular and complex type of molecular descriptors that represent the molecular structures and properties of molecules. These features are usually encoded as binary bit vectors, whose purpose is to reproduce (in several different ways), a characteristic pattern of the given molecule[12], [19, 20, 21, 22].

Chemoinformatics tools and data files formats

Chemoinformatics tools are quite valuable since they help to speed up and automate a lot of important tasks of the RDD process. In this project were only used tools to compute molecular descriptors and tools to convert among data formats.

Tools for the calculation of molecular descriptors can transform chemical information which is encoded into a symbolic representation of the numbers in the molecule, or receive files with the chemical information of molecules and calculate the molecular descriptors.

Currently, the available tools to calculate molecular descriptors and for data conversion are:

PaDEL Descriptor is an open source software used for the calculation of molecular descriptors and fingerprints. Through the chemical formula of a given substance, this software is able to translate it into mathematical values, which provide detailed information about the substance to be explored. It has the ability to calculate 1875 descriptors (1444 descriptors (1D, 2D) and 431 3D descriptors) and 12 types of fingerprints. It is a program developed using the Java[13] language and consists of a library and an interface component[14].

OpenBabel is a chemoinformatics tool designed to read various chemical data formats. It is an open source software. It can be used for several purposes, such as, filtering, conversion,

[12] "Descritores moleculares para aprendizagem automática".Available in http://joao.airesdesousa.com/agregacao/slides_2013/descritores_QSPR_slides.pdf, accessed last time in 24-11-2016

[13] Java is a programming language and computing platform.

[14] Padel-descriptor, 2014.Available in www.yapcwsoft.com/dd/padeldescriptor/, accessed last time in 25-10-2016

analysis and storage of molecular modeling data. This tool is commonly used for the proliferation of multiple formats of chemical files, since it has the capacity to convert about 110 formats[15], [23].

PowerMV is a software used for statistical analysis, molecular viewing, calculation of descriptors and similar researches. It supports files in SDF format and properties of molecules can be exported to Excel to generate custom reports[16].

MODEL (Molecular Descriptor Lab) is a software that allows the calculation of structural and physico-chemical properties of molecules from its 3D structure [24].

JOELib is a software used for the chemical conversion of file formats. It is a chemoinformatics library programmed in Java and has an interface with external programs[17].

Visual Molecular Dynamics (VMD) is a computer program used for molecular analysis and for the analysis of large biomolecular systems using 3D graphics and embedded scripts. This program allows a modeling and analysis of several biological systems, such as proteins. VMD is very easy to use and run on MacOS X, Unix and Windows [25]. There is no limit on number of molecules or atoms viewed, supports more than 60 molecular file formats and has extensive documental support[18].

These tools support files in various formats[19].

Structure Data Format (SDF) belongs to the chemical data file formats developed by Molecular Design Limited (MDL). It is a chemical file format used to represent chemical structures and has the interesting feature of allowing the inclusion of extra associated data. Due to its characteristics, it is the most widely used standard for importing and exporting information on chemical data[20].

Another format commonly used for molecule notation is the Simplified Molecular Input Line Entry Specification (SMILES). In this language there are two main types of symbols, atoms and bonds. This format has the advantage of presenting the information in an understandable and very compact way[21]. The SMARTS format is used to match chemical file substructures, using rules that are direct extensions of SMILES[22].

Chemical Markup Language (CML) acts as a support for chemical substances, such as molecules, and also for reactions, spectros, analytical data, among others. Manages molecular information using Extensible Markup Language (XML)[23] and Java languages. This format has the advantage

[15] Open babel:The open source chemistry toolbox, 2011.Available in http://openbabel.org/wiki/Main_Page, accessed last time in 12-10-2016

[16] Powermv.Available in www.niss.org/research/software/powermv, accessed last time in 25-10-2016

[17] Joelib/joelib2.Available in https://sourceforge.net/projects/joelib/, accessed last time in 25-10-2016

[18] VMD program.Available in www.ks.uiuc.edu/Research/vmd/, accessed last time in 20-01-2017

[19] There are, currently, around 100 file formats for chemoinformatics data.

[20] How to create SD/SDF files.Available in https://eurl-ecvam.jrc.ec.europa.eu/laboratories-research/predictive_toxicology/qsar_tools/qrf/How_to_create_SDF_files.pdf, accessed last time in 18-11-2016

[21] SMI file extension.Available in www.file-extensions.org/smi-file-extension-daylight-smiles-file, accessed last time in 18-11-2016

[22] SMARTS - A Language for Describing Molecular Patterns.Available in www.daylight.com/dayhtml/doc/theory/theory.smarts.html, accessed last time in 18-11-2016

[23] XML is a language that defines a set of rules for encoding documents in a format that can be understood by humans and machines.

of using XML portability to help interoperate documents between the various entities of interest[24].

International Chemical Identifier (InChI) is a text identifier for chemicals. It is used to standardize the coding of molecular information and to facilitate the search of this information, thus allowing an easier connection between the various data collections. This format was developed during a project of the International Union of Pure and Applied Chemistry (IUPAC)[25].

Web repositories

There are many web repositories with available databases storing drugs efficacy tests. Some of them allow public access, others have a cost.

Distributed Structure-Searchable Toxicity (DSSTox) is a project that belongs to the Computational Toxicology Research Program of the United States Environmental Protection Agency. It is a repository that provides a fairly complete chemical resource, which contains toxicity data present on certain molecules and chemical structures[26]. This database is free, so it can be consulted by the general public, thus promoting research and development. The format of this data is the SDF, already explained above[27].

US Food and Drug Administration (FDA) is a well-known regulatory agency of drugs that is responsible for ensuring the safety and efficacy of drugs in order to preserve the health of population. It has an open source database online[28].

Pre-clinical And Clinical Trials knowledge base on bioavailability (PACT-F) is a database structure-based and consists of the results of human clinical trials and pre-clinical animal tests. PACT-F exists since 2005 and is currently the largest bioavailability database in the world and contains approximately 8296 clinical trial registrations. These records are described in detail online and contain the chemical structure of the various compounds, thus allowing scientists to relate this structure to bioavailability. PACT-F contains all the important information for the development of predictive models that will allow to study the bioavailability of the new compounds. PACT-F is a database that allows the development of computational models, selection of candidate drugs based on previous assays, exploring the various factors that affect bioavailability, identifying some structural patterns that may influence bioavailability and drug optimization[29].

The National Library of Medicine (NLM) TOXNET is a free web repository, managed by the Toxicology and Environmental Health Information Program (TEHIP), which contains a group of databases related to chemicals, pharmaceuticals, general health and also toxicology. This repository can be used to search for chemical nomenclatures, chemicals, chemicals associated

[24]CML - Chemical Markup Language.Available in www.ch.ic.ac.uk/rzepa/cml/, accessed last time in 18-11-2016

[25]THE IUPAC INTERNATIONAL CHEMICAL IDENTIFIER (INCHI).Available in https://iupac.org/who-we-are/divisions/division-details/inchi/, accessed last time in 18-11-2016

[26]DSSTox.Available in https://datahub.io/dataset/dsstox, accessed last time in 30-11-2016

[27]DSSTox project.Available in www.epa.gov/chemical-research/distributed-structure-searchable-toxicity-dsstox- accessed last time in 30-11-2016

[28]U.S FOOD DRUG ADMINISTRATION.Available in www.fda.gov/default.htm, accessed last time in 30-11-2016

[29]PACTF.Available in www.pharmainformatic.com/html/pact-f.html, accessed last time in 30-11-2016

with diseases, toxic effects of some chemicals for the health of humans or animals, among others. On the web page is available the list of all the databases that constitute TOXNET, with a brief description of each one[30].

PharmacoKinetics Knowledge Base (PKKB) is a free web repository that contains a larger database extension with very relevant information concerning ADMET tests. This repository contains data from about 1685 drugs covering various test properties, such as, solubility, intestinal absorption, volume of distribution, urinary excretion, among others[31].

ADMEdata is a web repository containing data sets obtained from different studies. The complete database has 597 compounds and about 9500 data points that allow the creation of predictive ADME models. This repository has available data sets on tests of absorption, distribution and also metabolism[32].

Super Toxic is a web repository that contains information about known toxic compounds and uses that information to assess the property of unknown substances. Super Toxic database provides several search options by name and also by properties, such as, molecular weight or measured toxicity values. The use of these data enables the risk assessment of the new products to be minimized and their subsequent toxicity effects[33].

Cheminformatics is a web repository that contains links to several chemoinformatics programs as well as several data sets. All data sets are free for academic purposes. This web site contains data sets on the combination of more than one ADMET properties, toxicity, molecules that penetrate BBB, among others [34].

Sometimes, repositories only present the names of the molecules and it is necessary to get their structure. ChemSpider is a free chemical structure database that allows structure search access and has many data sources for more than 58 million structures[35].

2.3 Data Mining

In recent times, there has been an exponential growth in the information to which people have access through the Internet [26]. Most of the institutions (public and private), store computer data of their activities, creating large data repositories [27]. Companies have monitored the development of the technologies, and over the years have been collecting and storing a large amount of data continuously. In the 1990s, particular attention was paid to these stored data. It was realized that the data were being underutilized and that it could be an added value for the strategic positioning of the companies. However, the high volume of data stored in disparate structures has quickly

[30] TOXNET.Available in www.nlm.nih.gov/pubs/factsheets/toxnetfs.html, accessed last time in 30-11-2016

[31] PharmacoKinetics Knowledge Base.Available in http://cadd.suda.edu.cn/admet/v, accessed last time in 15-01-2017

[32] ADMEdata.Available in www.admedata.com/, accessed last time in 15-01-2017

[33] Super Toxic project.Available in http://bioinf-services.charite.de/supertoxic/index.php?site=home, accessed last time in 15-01-2017

[34] Cheminformatics.Available in http://cheminformatics.org/, accessed last time in 15-01-2017

[35] ChemSpider.Available in www.chemspider.com/, accessed last time in 21-01-2017

become overwhelming. Thus, the need to create databases and database management systems arose [26]. Today, society is commonly referred to as being in the "information age". With the great expansion of the use of new technologies, the markets demand professionals to be prepared for this evolution. So, there has been an increase in tools to organize and manage all the data, allowing the discovery of information in databases which apparently did not exist or were hidden [28]. Due to the great need for data analysis tools without limitations, researchers resort to ideas and methods developed in Machine Learning[36]. This search for new ideas led to the emergence of a new area of research, Knowledge Discovery in Databases (KDD), currently also known as Data Mining (DM) [3, 26].

Often, the concepts of DM and KDD are used as synonyms. Knowledge Discovery in Databases (KDD) is a non-trivial process of identifying, validating and recognizing data patterns that may provide valid information by generating useful and unexplored knowledge about a specific database. This process describes the extraction of knowledge from the data and generally consists of a set of various steps such as data cleansing, data integration, data selection, data transformation, data mining, evaluation of standards and finally, the evaluation of knowledge, as it is possible to see in Figure 2.5. Data Mining is a step of the KDD process, and is at the heart of the process of knowledge exploration. It is at this stage that the techniques and algorithms that will be used in the problem are applied, in order to extract data models [29].

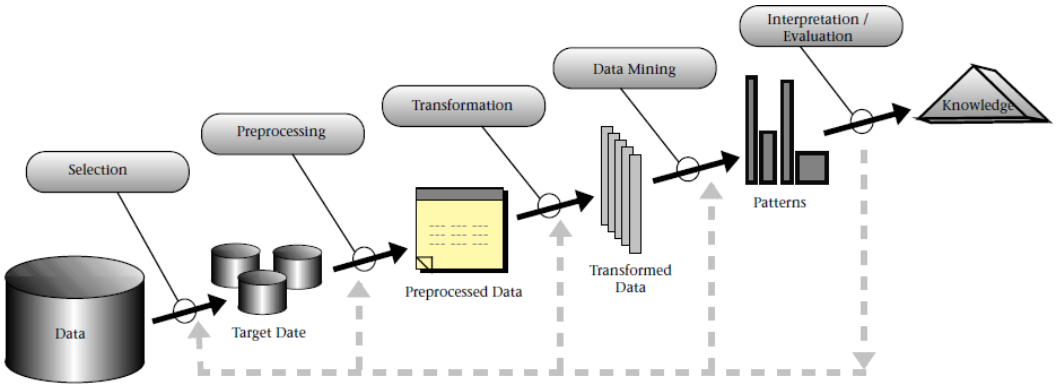


Figure 2.5: An overview of the steps that compose the KDD process. Extracted from [3].

Data Mining aim to extract useful regularities from long data archives in two possible ways, directly in the form of "knowledge" - characterizing the relations of the variables of interest, or indirectly - as functions that allow prediction, classification or representation of the regularities, in the distribution of the data [26]. Data Mining in hospital databases and public health information systems makes it possible to discover relationships so that a prediction of future trends can be made, allowing a better characterization of the patient. However, some health professionals have difficulty understanding the data. This can probably happen due to the high rate at which data is generated, which leads to an inability of the human being to extract and interpret this data [27].

[36]This area develops computational models to acquire knowledge based on previous facts.

Data Mining is one of the most promising technologies today [30]. It aims to find hidden relationships between data to allow predictions for the future. It is a process of extracting useful data from large volume databases, which means, it consists of the discovery of interesting knowledge, such as patterns, associations or anomalies [26].

Data Mining is considered as an extension of the traditional method of data analysis, since it consists of several analytical techniques. DM is used in many areas, from the most traditional to the most modern, due to the ease and speed of access to the data and the information needed. It has applications in the field of marketing, science, medicine, telecommunications, financial data analysis, retail industry, Adaptive Business Intelligence (ABI) among others. Thus, this may be the solution to the data analysis problems that many organizations face daily. However, this technology is still very recent, and so it is necessary to research to make it more efficient [31].

There are several processes that define and standardize the different phases and activities of Data Mining. The CRISP-DM (Cross-Industry Standard Process of Data Mining) [32] is the most used model due to the high information available in the literature and because it is currently considered the most accepted standard [30]. This is divided into 6 cyclic steps and the flow is not unidirectional as can be seen in Figure 2.6.

The phases of CRISP-DM are:

1. Business understanding - this step is important for understanding the different organizations that can use Data Mining, thus helping the next steps.

2. Data understanding - it is necessary to know the source from which the data came from, who collected them, what they mean, etc. Thus, once the objectives are defined, knowledge of the data is necessary.

3. Data preparation - data comes in various forms and formats. Due to the different possible sources, it is likely that the data is not prepared for the application of Data Mining methods. Data preparation can involve a number of activities, depending on the quality of the data.

4. Modelling - a model, in the area of Data Mining, is a computerized representation of real observations. In this step DM algorithms are applied based on the intended objectives.

5. Assessment - this is considered the critical stage of the Data Mining process. It is necessary the participation of experts and knowing of the business to do the evaluation. This phase is specific to help determine if the model is valuable and what can be done with it. To test the reliability of the models, some tests and validations must be performed. Indicators should also be obtained to assist in the analysis of results, such as confusion matrix, correction index, kappa statistic, mean absolute error, precision, F-measure, among others.

6. Deployment - in this step there is a need to organize and present knowledge in a way that the client can use. It involves several steps such as planning, monitoring and maintaining the

plan, producing the last report and reviewing the project. Usually, these steps are performed by the client[37] [30, 33].

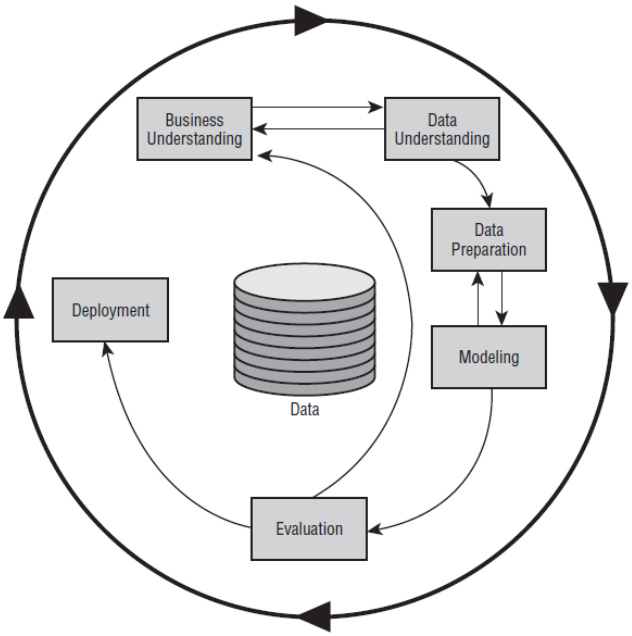


Figure 2.6: Phases of CRISP-DM process. Extracted from[37].

Data Mining tasks

Generally, Data Mining is classified according to its ability to perform different tasks [30]. These tasks are very distinct as there are many types of patterns in a massive database. Thus, different methods and techniques are needed to meet different types of patterns.

Depending on the patterns we are looking for, Data Mining tasks can be classified into classification, association rules discovery, clustering, outlier detection and regression [30, 34, 35].

Classification

Classification is one of the most common tasks and its final objective is to identify which class a particular record belongs to (through the analysis of previous records) [30]. It is a supervised learning method, that is, it uses sets that were previously classified to predict the classification of the new observation [36].

Classification defines a set of models that can be used to classify new objects. The models are constructed from a pre-analysis of the data set of a sample that possess objects classified in a correct form. These models can be represented in several forms, such as, Decision Trees, Random

[37] CRISP-DM.Available in www.explore-thinking.com/bring-order-into-the-social-media-monitoring-process/, accessed last time in 25-10-2016

Forest, Naivy Bayes, rule-based, k-Nearest Neighbors (k-NN), Support Vector Machine (SVM), among others [27, 30, 31, 34, 37].

Classification algorithms

Decision Trees: works as a tree-shaped flowchart, that is, it has the shape of an inverted tree that has its root at the top and is divided successively into several branches. The input attributes are represented by the inner nodes of the tree and each branch represents a possible result for this test. The lower nodes are called sheet nodes and represent the classes that the algorithm is attempting to predict[38].

This method is quite simple to interpret, it does not require configuration parameters, it also does not require much data preparation and it can deal with missing values. As disadvantage, this algorithm can originate super-complex trees, causing overtting.

All decision tree algorithms follow a set of steps:

- Selection of an attribute A in order to divide;

- Subdivision of attribute A in disjoint subsets;

- A tree is returned with attribute A as a root and with a set of branches on the lower level. Each branch will have a descending sub-tree;

There are several algorithms designed for the construction of decision trees: ID3,C4.5,CART and CHAID. Different algorithms are distinguished by the criterion that decides which next node to be explored and the type of test performed at each node inside the tree [35, 38].

In the classification experiments performed it was used the CART algorithm.

For the CART algorithm, the selection criterion used to choose the A node to be explored is Gini Index or Gini impurity metric. The node having the least impurity is chosen and the impurity is calculated by the formula:

$$GiniIndex(f) = \sum_{i=1}^{J} fi(1 - fi) \tag{2.1}$$

where, $\boldsymbol{J}$ represents the number of classes and $\boldsymbol{fi}$ corresponds to the fraction of items labeled with class i.

As a general rule, the decision tree algorithm ends its recursion when all data examples have the same label value. However, there are also other conditions that may lead recursion to end: when there are less than a number of instances or examples in the current sub tree (this problem can be adjusted through the parameter "minimum size for split"), when no attribute achieves a given minimum gain, relative to the existing tree before that division (a

[38] Classification And Regression Trees for Machine Learning, by Jason Brownlee, 2016. Available in http://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/, accessed last time in 21-05-2017

solution for this problem can be set via the "minimal gain parameter"), and lastly, when the maximum depth is reached (this can be adjusted through the "maximal depth parameter") [35, 38].

Random Forest: corresponds to a large number of decision trees. Each observation is fed into each decision tree. The number of decision trees generated is specied by the programmer. The resulting model is a voting model among all the generated decision trees. For example, a certain attribute "x" belongs to class "1" if was predicted by most trees generated as belonging to class "1" [35].

Each decision tree is generated in the same way as the decision trees specified in the previous algorithm. The only difference is that for each division, random forest selects the node from a random set of nodes, instead of being selected from among all available nodes to be branched (decision trees algorithm) [39].

Naive Bayes: is an algorithm based on the Bayes' theorem. This is a supervised learning algorithm that assumes that the presence (or absence) of a given attribute is not related to the presence (or absence) of any other attribute [40], [39].

Given a role = label 'C' attribute of classes c1, c2, ..., cn and an attribute vector 'a' corresponding to all other attributes, the conditional probability of a class ci can be defined as:

$$P(C = ci|a) = \frac{P(C = ci) * P(a|C = ci)}{P(a)} \tag{2.2}$$

Rule-based classification: has the structure -if- condition -then- conclusion and is sometimes retrieved from a decision tree [27, 40]

k-NN: corresponds to a test of structural similarity between test and training objects. The lower the distance, the more similar these compounds will be. This technique uses a space representation of instances in order to classify different instances of unknown classes. More specifically, k-NN has its training elements in space, each associated with the class to which it belongs. In order to classify the test elements, it is necessary to verify the votes of the nearest neighbors, and the new element is classified by the most frequent answers. If k = 1, the element is classified with the class corresponding to its nearest neighbor (and so on for the various values of k) [35].

[39] Random Forests Algorithm, Michael Walker, 2013. Available in www.datasciencecentral.com/profiles/blogs/random-forests-algorithm, accessed last time in 21-05-2017

[40] Naive Bayesian, Dr. Saed Sayad. Available in www.saedsayad.com/naive_bayesian.htm, accessed last time in 21-05-2017

SVM: is a relatively recent technique but has been widely recognized for its results. This is a supervised machine learning algorithm that is used both in classication and regression [41].

Given a set of training examples, each marked as belonging to one of two classes, the SVM algorithm constructs a model that represents a number of examples with points in space. These points are mapped so that the examples in each category are divided by a straight line. Examples that belong to different classes should be represented as far as possible.

The new examples, present in the test file, are then mapped in the same space and predicted as belonging to one of the two classes, according to the side of the line in which they are placed. When data are not linearly separated, it is necessary for the SVM algorithm to resort to the kernel function.

SVM's main problem is the time it takes to train a model [22], [30].

Association rules discovery

Association rules discovery is a task responsible for most of the solutions that are used for pattern discovery. The algorithms of association rules discovery aim to discover rules to quantify the correlation between two or more attributes, thus allowing the understanding of the new models. This correlation is commonly referred to as the association rule. This association rule reveals whether the appearance of a set of objects in a database, is closely related to the appearance of another set [27, 29]

Clustering

Clustering is a task that does not depend on the existence of pre-sorted data. Therefore, this is an unsupervised learning technique that learns from observation and not from examples. A successful cluster produces high-quality clusters to ensure that objects within a cluster are very similar to each other, but very different from objects in other clusters. Once clusters are formed, the objects are labelled and their common characteristics are summarized to form the class description. This task is similar to classification, with the difference that in the classification, classes are pre-defined and in clustering, classes are dynamically created based on the similarity between the elements. Clustering can be used in several areas, such as, image processing, data analysis, pattern recognition, among others. This task has some algorithms, the most common are k-Means and k-Medoids [30, 31, 37, 40, 41].

Clustering algorithms

K-Means: is a method in which, in a data set, the algorithm selects randomly k records, each representing a cluster. Then, for each record that remains, the similarity between it and the

[41] Understanding Support Vector Machine algorithm from examples (along with code), by SUNIL RAY , 2015. Available in www.analyticsvidhya.com/blog/2015/10/understaing-support-vector-machine-example-code/, accessed last time in 21-05-2017

center of each cluster is calculated, and the object is placed in the cluster that has the greatest similarity.

k-Medoids: is an algorithm that appears as a variation of the previous method. The main difference is that in this technique, instead of calculating the center of the cluster and to use it as reference, the most central object of the cluster is used.

Partitioning and hierarchical methods generate clusters that are usually spherical. However, in some cases this distribution may be denser and the results are not the most satisfactory. Thus, density-based methods are emerging and allow for better results . Given an aleatory set of points in space, it groups together points with many nearby neighbors, marking as outliers the points that lie alone in low-density regions (whose nearest neighbors are too far away). An example of an algorithm of this type is Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [30].

Outliers

Outliers are exceptions (data elements) that cannot be grouped into classes or clusters. Although they may be considered as noise and discarded in some applications, in others may bring important information and their analysis is quite significant [30, 31].

Regression

Regression is a supervised learning task, often compared to classification (classification groups finite sets of discrete values and regression predicts continuous values). It is also used to find out which independent variables are related to the dependent variable, and to understand this relationship. Thus, regression consists in estimating the value of a given variable, by analyzing other attribute values in the same item. The model obtained by regression can be represented in the form of linear regression, logistic regression, among others [30, 31, 42]..

Regression algorithm

Linear regression: is used to estimate a relationship between two variables through the mathematical formula: Y = mX + b. For example, in a graph with a Y and an X axis, the relationship between X and Y is a straight line [30, 31, 42].

Other algorithms: some of classification algorithms, such as Decision Trees, Random Forest, SVM and k-NN, are also regression algorithms.

Data Mining tools

In order to facilitate the process of Data Mining, several useful tools have been developed. The most popular ones are described throughout this section.

Waikato Environment for Knowledge Analysis (WEKA): is a set of machine learning algorithms for DM that was implemented in 1997 in New Zealand. It is a very popular open source tool developed in Java (which allows it to be used on several platforms). This software provides the necessary functionality to perform various tasks such as association rules discovery, regression, clustering and has several algorithms inside each task[42],[43]. The WEKA system has a simple graphical interface and its biggest advantage is the fact that it allows comparing the performance of different techniques, verifying which one has the lower error rate [30, 43, 44, 45].

KNIME: is a tool that allows doing the analysis of patterns, predicting potential results in the data and also discovering trends. This software is free and its graphical interface allows the user to gather several nodes to perform the pre-processing, modelling, analysis and visualization of data[44], [46].

RapidMiner: is an open source tool that was developed in 2001. This platform provides an integrated environment for machine learning, DM, between other areas. RapidMiner has algorithms to construct models for predictive analysis and also enables the graphical visualization of the results. This tool has the advantage of allowing that complex analyzes and resulting forecasts can be integrated directly into the customer's infrastructure, and also has the advantage that it is rarely necessary to write codes, which makes it a very intuitive tool [47].

R tool: in addition to being a scripting language, is also an integrated development environment for statistical and graphical use. This is widely used for data analysis and provides a wide range of statistical and graphical techniques (such as linear and nonlinear modeling, classification, clustering, among others). An important feature of this tool is the ease with which mathematical symbols and formulas can be manipulated whenever necessary. However, it has the disadvantage of its unfriendly interface[45].

Evaluation metrics: classification

Before using a model, it is necessary to evaluate it by testing its behaviour with real data and then doing its validation using good evaluation practices. For this, different metrics have been proposed, each of which evaluates different characteristics of the data classifier [4, 48]. The main characteristic of a classification system is that it can predict which class an object belongs to,

[42]"Mineração de dados com WEKA, Parte 1: Introdução e regressão".Available in www.ibm.com/developerworks/br/opensource/library/os-weka1/, accessed last time in 25-10-2016

[43]Weka 3: Data Mining Software in Java.Available in www.cs.waikato.ac.nz/ml/weka/, accessed last time in 25-10-2016

[44]Examining the KNIME open source data analytics platform.Available in http://searchbusinessanalytics.techtarget.com/feature/Examining-the-KNIME-open-source-data-analytics-pla accessed last time in 25-10-2016

[45]R language.Available in www.r-project.org/about.html, accessed last time in 25-10-2016

otherwise there is an error[46],[47]. The performance of a classification system is calculated by the ratio of the number of errors found to the total number of instances tested. Metrics play a crucial role in selecting the best classifier during the training[48], [4].

Data classification problems may be divided into binaries, multiclass, and multi-labelled classification.

Accuracy, Error rate, Precision, Recall, F-measure and AUC are the metrics addressed in this report[49], [4].

Accuracy: is one of the most used metrics to evaluate the generalization of classifiers, whether in binary or multiclass classification problems. It is based on the proportion of correct predictions for all instances, which means, its purpose is to estimate how often the classifier makes the correct predictions. The main advantage of this metric is that it is easy to calculate and easy to understand by users. The main disadvantages are that it does not differentiate classes (positive or negative) and leads to poor values of discrimination (less rigor in choosing the right solution).

Accuracy is defined as:

$$\frac{tp+tn}{tp+fp+tn+fn} \tag{2.3}$$

where, ***tp*** - true positive for ***C***; ***tn*** - true negative for ***C***; ***fp*** - false positive for ***C***; ***fn*** – false negative for ***C***.

Error rate: like accuracy, also belongs to the most commonly used metrics and they complement each other. This represents the proportion of errors committed in relation to the total set of instances, and this percentage is used to evaluate the solution produced. Its main advantage is the ease of performing the calculation. Error rate is defined as:

$$\frac{fp+fn}{tp+fp+tn+fn} \tag{2.4}$$

where, ***tp*** - true positive for ***C***; ***tn*** - true negative for ***C***; ***fp*** - false positive for ***C***; ***fn*** – false negative for ***C***.

Precision: is used to identify which predictive items are true and so gives the ratio of 'tp' to all positive results. The precision and recall metrics have only one evaluation task (positive or

[46]Testing and Validation (Data Mining).Available in https://msdn.microsoft.com/en-us/library/ms174493.aspx), accessed last time in 03-01-2017

[47]Turi Machine Learning Platform User Guide.Available in https://turi.com/learn/userguide/evaluation/classification.html), accessed last time in 03-01-2017

[48]Classification Accuracy is Not Enough: More Performance Measures You Can Use, Jason Brownlee, 2014.Available in http://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/), accessed last time in 03-01-2017

[49]Model Evaluation.Available in www.saedsayad.com/model_evaluation.htm), accessed last time in 03-01-2017

negative class), so do not have the possibility to choose the optimal solution. Usually, these techniques are complementary and used together. Precision can be defined as:

$$\frac{tp}{tp+fp} \tag{2.5}$$

where, ***tp*** - true positive for ***C***; ***fp*** - false positive for ***C***.

Recall: corresponds to the proportion of positive cases that are correctly identified, and can be defined as:

$$\frac{tp}{tp+tn} \tag{2.6}$$

where, ***tp*** - true positive for ***C***; ***tn*** - true negative for ***C***.

Area Under the Curve (AUC) and F-measure: are metrics also used for discrimination in binary classification problems. The main advantage of AUC is its excellent effectiveness. However, it has the disadvantage of high computational cost, specially when it is used in multi-class classification problems. The AUC values represent the performance of a classifier and the value 1 corresponds to a perfect classifier.

The AUC value can be calculated by:

$$\frac{Sp - n_p(n_n+1)/2}{n_p * n_n} \tag{2.7}$$

where, ***Sp*** represents the sum of the all positive examples ranked, while n_p and n_n means the number of positive and negative examples respectively.

F-measure is defined as below:

$$\frac{2*p*r}{p+r} \tag{2.8}$$

where, ***p*** - precision; ***r*** - recall.

Table 2.1 below presents a summary of the definition of the most used evaluation metrics.

Table 2.1: Metrics for classification evaluations. Adapted from [4].

Metrics	Evaluation focus
Accuracy	Measures the ratio of correct predictions over the total number of instances evaluated;
Error Rate	Measures the ratio of incorrect predictions over the total number of instances evaluated;
Precision	Measures the positive patterns that are correctly predicted from the total predicted patterns in a positive class;
Recall	Measures the fraction of positive patterns that are correctly classified;
F-Measure	Represents the harmonic mean between recall and precision values.

The degree of reliability of each metric is estimated with the confusion matrix [48, 49, 50, 51, 52]. Confusion matrix allows to define the four quantities already mentioned in the formulas of each metric:

True Positives (tp): corresponds to the number of positive molecules predicted correctly;

True Negatives (tn): as the name implies, it corresponds to the number of negative molecules predicted correctly;

False Positives (fp): corresponds to the number of negative molecules that are predicted to be positive;

False Negatives (fn): corresponds to the number of positive molecules that are predicted to be negative [51].

The confusion matrix simplifies the understanding of these concepts and is represented in the Table 2.2.

Table 2.2: Example of a confusion matrix.

		Real	
		Positive	Negative
Predicted	Positive	tp	fp
	Negative	fn	tn

Evaluation methodologies: classification

In the process of developing a model, the evaluation phase is a crucial process and must be done rigorously. This evaluation allows the verification of the behaviour of the models, and allows their choice according to what has the best characteristics.

Cross Validation is one of the methods of evaluating models in Data Mining[50]. It consists on the evaluation of the general capacity of a model. This technique has several types:

Hold-Out method: consists of dividing the data set into two main subsets, training set and test set (sometimes it is also divided into three subsets, adding the validation set). Figure 2.7 presents an example of this method. The training set is used in the building of predictive models and in the train of the classifier. The validation set is used to evaluate the performance of the model produced in the training phase, and may not occur in all modeling algorithms. The test set enables to estimate the error rate of the classifier and also the future performance of the model. The main advantage of this technique is that it presents independent training and testing. On the other hand, it has the disadvantage of not using all available data (performance evaluation may be very different) and depending on the way the training / testing division is done[51].

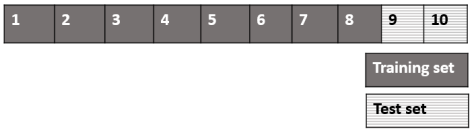


Figure 2.7: An example of the Hold-Out method using 80% of the examples for training and 20% for testing.

k-fold Cross Validation: is the most common and consists of dividing the data set into k parts of approximately equal size. The model is built in all folds, but one is left out and is used to test the prepared model while the others are used as training set. The main advantage of this approach is the displaying of an accurate estimation of the performance of the model and the disadvantages are overlapped training data, small samples of performance estimation and underestimated performance variance[52], [53].

In Figure 2.8 it is possible to see an example of k-fold Cross-Validation technique, with k=2.

In this example, data is divided into 2 and therefore, two models are built. For each model half of the examples is used for training and the other half of the examples is used for testing.

Leave-One-Out Cross Validation: is a specific case of k-fold Cross Validation when k is equal to the sample size (N), that is, equal to the number of instances in the data. An example of this Cross-Validation method can be seen in Figure 2.9. The advantage of using this

[50]Cross Validated.Available in http://stats.stackexchange.com/questions/103459/how-do-i-know-which-method-of-cross-validation-is-best, accessed last time in 03-01-2017

[51]Hold-Out method.Available in scott.fortmann-roe.com/docs/MeasuringError.html, accessed last time in 03-01-2017

[52]Model evaluation, model selection, and algorithm selection in machine learning, Sebastian Raschka, 2016.Available in https://sebastianraschka.com/blog/2016/model-evaluation-selection-part3.html, accessed last time in 03-01-2017

[53]k-fold.Available in https://pt.slideshare.net/hafidztio/l1penalization?smtNoRedir=1, accessed last time in 03-01-2017

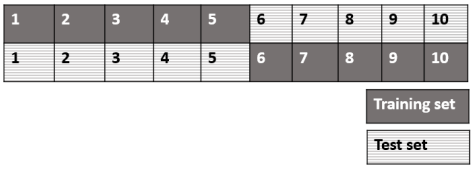


Figure 2.8: An example of the 2-fold Cross Validation method.

approach is that it provides an unbiased performance estimate. On the other hand has the disadvantage of having a high computational cost and not allowing stratification[54], [55].

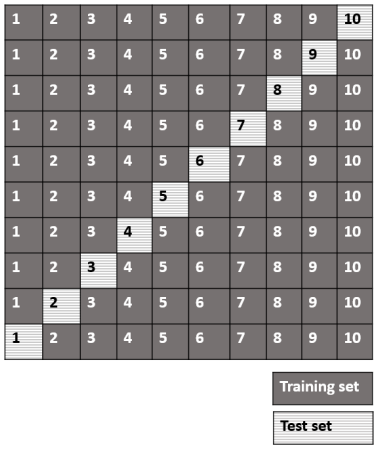


Figure 2.9: An example of the specific case of Leave-One-Out Cross Validation method.

As it is possible to verify, this method uses only one of its examples for testing, and all other examples as training.

Evaluation metrics: regression

There are several parameters that are considered when it comes to evaluate the regression models. Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are the metrics addressed in this report.

Root Mean Squared Error: is a measure that provides the deviation between the observed values and the predicted activity values. RMSE is often used and it makes an excellent general purpose error metric for numerical predictions.

Root Mean Squared Error is mathematically defined as:

[54] How to Evaluate Machine Learning Algorithms, Jason Brownlee, 2013.Available in http://machinelearningmastery.com/how-to-evaluate-machine-learning-algorithms/, accessed last time in 03-01-2017

[55] Leave-One-Out method.Available in www.projectrhea.org/rhea/index.php/Leave-one-out_Cross_Validation_Old_Kiwi, accessed last time in 03-01-2017

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{2.9}$$

where, $\boldsymbol{y}$ represents the values of experimental activity, $\hat{\mathbf{y}}$ symbolizes the expected activity values and $\boldsymbol{N}$ represents the number of molecules.

Mean Absolute Error: represents the information about the average of the deviations obtained in predicting activity.

$$\frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{2.10}$$

where, $\boldsymbol{y}$ represents the values of experimental activity, $\hat{\mathbf{y}}$ symbolizes the expected activity values and $\boldsymbol{N}$ represents the number of molecules [22, 53].

2.4 Ontologies

As the volume of data currently available is growing at a very large pace, it is necessary to use techniques of data processing and organization. The same concepts and definitions can be interpreted in different ways by the community. Thus, the need to create ontologies to standardize concepts and create norms arose. Ontology is an information organization technique that has its structure based on the description of concepts and semantic relationships between them[56],[57], [54, 55, 56].

The main objective of ontologies is to facilitate the sharing and reuse of information. For this to be possible, it is necessary that the concepts present in the ontologies have a formal specification.

In 1997, Borst [57] described that "ontologies are defined as a formal specification of shared conceptualization". From this definition, it is possible to understand that ontologies are very important in the various systems that aim to look for information coming from diverse communities, such as the information contained online.

There are several areas in which ontologies can be used, such as information retrieval, natural language processing, knowledge management, among others.

Their use brings some advantages such as to enable people to communicate about a particular knowledge, to allow the expulsion of contradictions and inconsistencies in the representation of knowledge and also to enable its developers to reuse knowledge bases, due to the formation of a consensus vocabulary. In spite of the various applications and advantages of ontologies, these still present some problems as the choice of the appropriate ontology for all the individuals or groups related to some specific domain, there can be problems related to the ontology library, since each one of ontologies could have been developed in a different context and there is also the problem of the lack of papers describing the methodologies for its development [57, 58, 59, 60].

[56] Ontologies and Semantic Web.Available in www.obitko.com/tutorials/ontologies-semantic-web/what-is-ontology.html), accessed last time in 09-01-2017

[57] Ontology, 2009.Available in http://tomgruber.org/writing/ontology-definition-2007.htm), accessed last time in 09-01-2017

Chemical ontologies

In the field of biology and chemistry, the amount of information available has been increasing exponentially. Thus, a more efficient organization of this data is necessary through the filtration process and posterior hierarchization of the concepts by computational methods. Thus, ontologies have been used to facilitate the organization of large volumes of data and allow a concise hierarchical classification [57, 58].

Chemical Entities of Biological Interest (ChEBI) is an ontology that constitutes a database, and that can be used for several attempts. It is available online for free[58] and aims at grouping chemical entities, such as, atoms and molecules. This ontology creates a hierarchical classification of molecules based on their structural chemical characteristics, biological functions and system interactions [57].

Hierarchical classification of molecules has several advantages, such as, allowing a compact representation of generalized knowledge at the highest level (which facilitates user understanding) and also plays an important role in useful predictions and predicting properties of new entities [57, 58].

The use of ontologies in the chemoinformatics domain is very relevant because it allows a better interpretation of biological mechanisms, which allows the detection of bioactivity patterns associated with the chemical structure of the molecules[59], [61, 62, 63].

Encoding ontologies

An ontology language is a formal language that is used to encode / save the ontologies. Currently there is a big diversity of languages to formalise ontologies. In this report are addressed only the most common[60].

The languages of ontologies can be classified according to sintax and logic, and the most common of the two classes are Knowledge Interchange Format (KIF), Developing Ontology-Grounded Methods and Applications (DOGMA), Resource Description Framework (RDF) and the Web Ontology Language (OWL)[60], [64].

KIF: provides the definition of objects, relationships, functions and it is easy to understand the logic. The meaning of the expressions can be understood without needing the interpreter to manipulate these expressions [64].

DOGMA: is another language that is not restricted to a particular language. It has some properties that distinguish it from other languages. It differs in its basis for the representations of knowledge and in its specific duality in the interpretation between the level of language and the conceptual level [64].

[58] www.ebi.ac.uk/chebi/

[59] Ontology, 2009.Available in http://tomgruber.org/writing/ontology-definition-2007.htm), accessed last time in 09-01-2017

[60] Ontology language.Available in https://en.wikipedia.org/wiki/Ontology_language, accessed last time in 21-01-2017

RDF: is used for conceptual description or modelling of information implemented in web resources. It uses various syntax notations and serialization data formats [64].

OWL: is the most known language and can include data as an interpretation of individuals and a set of properties that relate them to each other. This language is used for the publication and sharing of ontologies in the World Wide Web (WWW).

OWL was developed with the purpose of extending the existing vocabulary in the RDF language (previously explained).

World Wide Web Consortium (W3C)[61],[62]assumes three variants of this language: OWL Lite, OWL DL and OWL Full, according to the levels of expressiveness. It is considered an essential technology for the future implementation of the semantic web and has therefore been used to search tools, formal fundamentals and language extensions [64].

2.5 Related work

When it comes to starting a project, it is extremely important to review the related work. It is important to know what has already been done, what methods and tools have been used, the problems faced, the progress made, the results achieved, what can be improved and what has not yet been done but can lead to great progress in the area. After the research, it was verified that the Data Mining algorithms have already been used in several situations for prediction studies.

Predictive studies

Prevention of cancer is a public health issue of unquestionable importance. The US National Toxicology Program (NTP) conducts chemical rodent bioassays to help identify substances that may have harmful effects in humans. However, these trials are costly and result in a time-consuming process, which has led to an urgent need for the construction of models that propose molecular mechanisms for various health problems such as carcinogenesis, mutagenesis and toxicity testing [65, 66, 67].

Inductive Logic Programming (ILP) [63] tools were used for that predictive studies. They produce rules that are more concise and accurate than attribute-based algorithms. Most of these studies used the PROGOL ILP tool [65, 66, 67].

iLogCHEM

Another related work is iLogCHEM tool. iLogCHEM is an application that was designed to predict a property of interest through the structural information in a set of drugs [68]. Different compounds having similar structures usually exhibit the same pharmacological activity. The

[61] Is the main organization of standardization of the World Wide Web.

[62] www.w3.org/Consortium/

[63] Inductive Logic Programming is an area of Artificial Intelligence (AI) that studies the construction of theories from examples and prior knowledge. This is at the intersection of machine learning and logic programming.

Structure-Activity Relation (SAR) studies the path to make the drug more active through the analysis of structural relationships.

The iLogCHEM system aims to provide the analysis of molecules, through Data Mining tools, in an easy and transparent way for the user [69]. This system relies on three basic principles: only refines fragments that appear in the database, filter duplicates and perform efficient homomorphism tests.

Initially the system receives a data set from molecules already tested and identifies a set of patterns that are characteristic in the molecules that have activity.

The main problem in the construction of systems is the description of molecules. However, finding a discriminating component often reduces to the problem of finding a Common Maximum Substructure (CMS), that is, finding a pattern with the maximum number of atoms and bonds that is present in several molecules in the database introduced. The first approaches to look for common subframes are based on ILP, that uses first-order logic and has a powerful representation language, but is inefficient when using large databases. In this system the Aleph ILP tool was used with atoms and bonds as background knowledge [69].

Although its representation is not as specialized as other systems, the iLogCHEM tool has several advantages. An advantage is that the system can use the high number of search algorithms that are implemented in ILP [68]. An important feature of the iLogCHEM tool is the possibility of adding extra information, which will facilitate the search process of the appropriate model. Thus, the user can choose which chemical structures he wants to detect in the molecules and then facilitate the understanding of the models. This important feature is achieved through the ability of ILP to accept background knowledge for rule discovery. In this system, the models are presented to the user through the visualization of the molecule and the pattern, using the Visual Molecular Dynamics (VMD) tool [25, 68].

The structure of the molecules is strictly related to their therapeutic effect. However, sometimes the drugs have the desired activity but not the required amount and other times lead to adverse effects. These problems can be solved in the RDD. So, the drug design process starts with a molecule that already has some activity, and then, are made modifications to improve it. However, there is the disadvantage of obtaining a large set of similar molecules. Thus, iLogCHEM has a solution. This system uses Tanimoto's distance (the user can choose the limit value for the distance used) to filter the pairs of similar molecules and retain only the most representative ones [68].

Other related work included an extension to the iLogChem tool. As mentioned above, this tool builds models using Data Mining algorithms that allow to determine whether a molecule is active or not. The already tested molecules are explored and patterns are found within them that allow inferred about its activity, both molecules and patterns can be visualized graphically and the specialist can also introduce restrictions on the creation of new rules.

This extension allows the inserted data set to have more than one format and not only provides the activity or inactivity of the molecule but also allows the created models to predict the ADMET

tests. In addition to the ILP system previously used by iLogChem, this extension uses Weka, that is another Data Mining tool [68, 70].

All the investigations and works carried out are important for those who have an interest in the area. It is important to make tests, use different methods and tools or continue studies already done, to achieve the objectives and bring the best living and health conditions to the humanity [68].

2.6 Chapter summary

In this chapter, it were introduced the basics concepts necessary to understand the project.

It was explained the RDD process that allowed the understanding of the problems of cost and time that are associated to this process, and that are the object of study of the work.

The concept of Data Mining was addressed, as well as its tasks, tools and other properties, which lead, after a detailed analysis, to the choice of the most appropriate tool to use in the work.

The methodologies and metrics to evaluate the experiments that were used in the study were established.

The most relevant web repositories for the study were identified.

Ontologies and their languages were described in the last part of the chapter. It was possible to understand their importance for the organization of large amounts of data, to standardize concepts that lead to a more precise sharing of knowledge and to enrich data sets.

It was also mentioned some related work to the subject of study.

Chapter 3

Experimental evaluation

In this chapter we first make a contextualization of the work and then we describe the steps involved in data preparation. The second part is divided in two sections: Toxicity experiments and Blood-Brain Barrier (BBB) penetration experiments. These two sections include a description of the data sets, the data pre-processing and the experiences performed.

3.1 Work plan contextualization

The dissertation work corresponded to a set of Data Mining experiments in order to answer the two research questions stated in Chapter 1.

Figure 3.1 presents the general block diagram of the tasks that were done in the dissertation work.

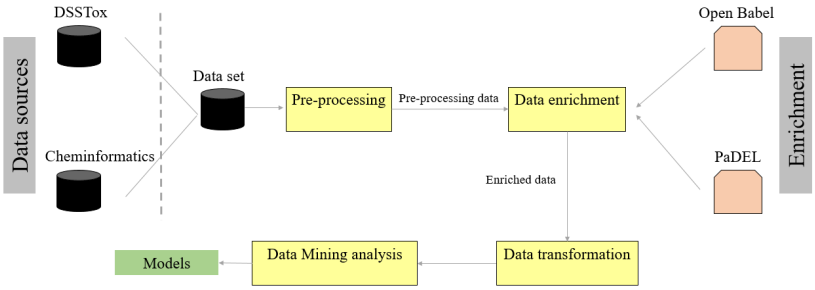


Figure 3.1: Brief explanation of the work that was developed.

As previously mentioned, the main objective of the work was to use Data Mining tools in order to improve ADMET tests.

The first step of the work was to choose a set of molecules previously tested in animals, which were downloaded from web repositories (in this case, DSSTox and Cheminformatics).

The following actions took place in the pre-processing stage. The data were prepared in different ways to obtain an adequate data set.

After pre-processing, the data undergoes an enrichment stage where added information was associated with each molecule. This extra information was computed using Open Babel and PaDEL tools.

The Data Mining tool chosen to carry out the experiments was RapidMiner due to its simple interface. We have used the educational version, since it allows free and unrestricted number of lines in the data sets (only for students and teachers).

As this work is prospective in nature, we have investigated the quality of several algorithms to assess which provided the best results.

For classification studies we have used: SVM, k-NN, Decision Trees with CART and Random Forest with CART algorithms.

In the regression study we have experimented Linear regression, SVM and k-NN algorithms.

3.2 Data preparation

As mentioned before, nowadays there is a large amount of data available and accessible. Data Mining is one of the approaches that can be used to explore this data. For this exploration, it is advisable to carry out some steps in order to prepare data, such as:

1. Data cleansing: when it comes to large databases, it is normal for inconsistent, incomplete or erroneous data to exist. For this reason, it is necessary to treat this data in order to avoid its influence on the results;

2. Data integration: since the data can have different origins, it is necessary to integrate it to avoid some problems, such as identification problems (the same concept can be designated differently), and thus to have a repository unique and consistent;

3. Data transformation: it depends on the desired objectives, some algorithms work only with categorical values and others with numerical values, it may be necessary to normalize the data (size the attributes on the same scale), among other transformations;

4. Data reduction: this phase aims to reduce the volume of data while maintaining the representativeness of the original data. It may include the elimination of attributes that are not of interest to the study that is being performed, leading to greater efficiency of Data Mining algorithms [71].

3.3 Cases studies

3.3.1 Toxicity Experiments

The first approach investigated was with toxicity data. Several experiments were made with these data, but the general plan of the study is represented in Table 3.1.

Experiments with toxicity data were performed with the original data and also with tanimoto-filtered data. On the right side of the Table 3.1 the scheme in which the original toxicity data was used is shown. On the left side one can see the scheme in which the data was previously filtered using a tanimoto coefficient.

Tanimoto coefficient filters similar pairs of molecules and retains only the most representative one. This coefficient was used to test if there was an improvement in the results when removing pairs of molecules with similar structure.

Table 3.1: Methodology for toxicity experiments.

ADMETDataSets = {Toxicity} DMalgorithms = {SVM & k-NN & Decision trees & Random forest} **Forall** DS in **ADMETDataSets** Enrich DS with Molecular Descriptors **Forall** Alg in **DMalgorithms** build and evaluate models **EndFor** **EndFor**	ADMETDataSets = {Toxicity} DMalgorithms = {SVM & k-NN & Decision trees & Random forest} **Forall** DS in **ADMETDataSets** Filter molecules by Tanimoto Enrich DS with Molecular Descriptors **Forall** Alg in **DMalgorithms** build and evaluate models **EndFor** **EndFor**

Data set

Distributed Structure-Searchable Toxicity (DSSTox) is a project that researches information about the toxicity of some molecules and chemical structures (already explained in section 2.2).

One of the data sets used for toxicity experiments was Carcinogenic Potency Database Summary (CPDBAS)[1]. Carcinogenic Potency corresponds to the level of toxicity present in molecules related to the chances of cancer.

The other data set used was the EPA Fathead Minnow (EPAFHM)[1]. U.S. EPA Mid-continental Ecology Division (MED) generated this database with the aim of developing a system to predict acute toxicity from the chemical structure based on the mode of action.

Data pre-processing

The original files were in SDF format, so the first step of the experiment consisted of using the OpenBabel tool to obtain the SMILES format of the molecules. Figure 3.2 represents the graphical interface of the OpenBabel software.

In order to apply a classifier, we need to define / compute the attributes to enrich the data set. This was enriched with molecular descriptors. As previously mentioned, these descriptors constitute a set of attributes that characterize the molecules.

The set of molecular descriptors was calculated using the PaDEL Descriptors software. As already mentioned in section 2.2, this platform contains a total of 1875 descriptors (1444 1D and 2D descriptors, and 431 3D descriptors). Figure 3.3 represents the graphical interface of the PaDEL software.

[1] www.epa.gov/chemical-research/distributed-structure-searchable-toxicity-dsstox-database

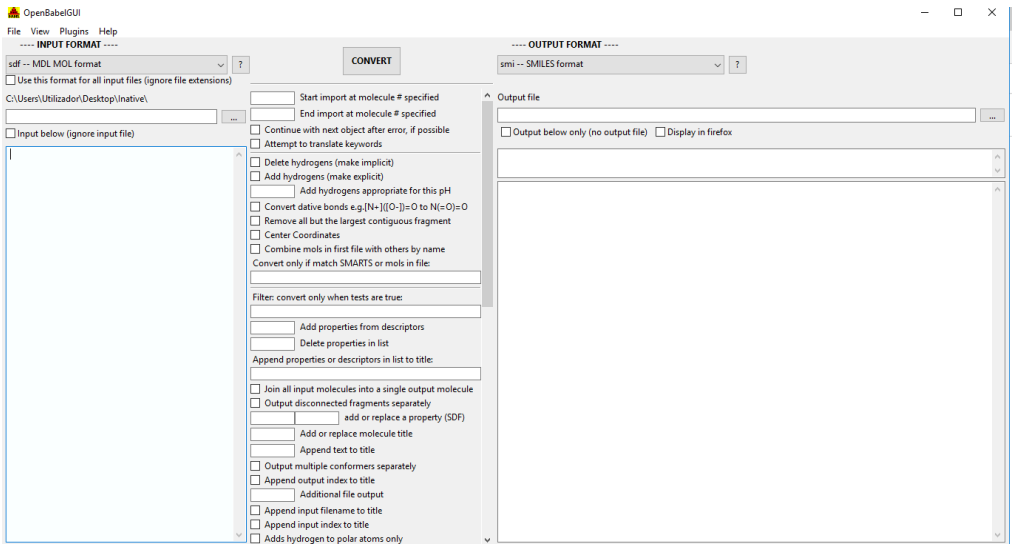


Figure 3.2: Representation of the graphic interface of the OpenBabel software.

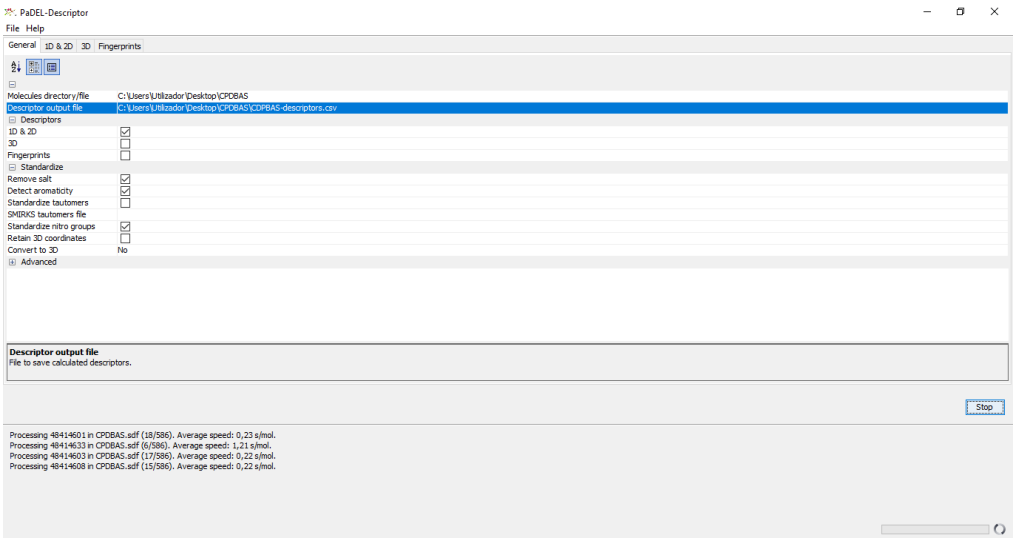


Figure 3.3: Representation of the graphic interface of the PaDEL software.

After calculating the 1D and 2D descriptors, the CPDBAS data set contained information on 1230 (586 molecules classified as 'active' and 644 classified as 'inactive) compounds and 1444 attributes, and EPAFHM data set contained information on 617 (580 molecules classified as 'active', 34 classified as 'inactive' and 3 had inconclusive results) compounds and 1444 attributes.

Some attributes representing molecular descriptors were removed because they contained a large number of missing values (they were not applicable to all molecules).

After removing missing values, CPDBAS data set had 1230 examples and 506 attributes, and EPAFHM data set had 617 molecules and 1212 attributes.

In EPAFHM data set we have also removed the molecules with inconclusive results, and therefore, the final data set contained 614 molecules.

The main reasons for the reduction of attributes are the decrease of the training time of each algorithm, and also an overall improvement of the model. Infrequent attributes can lead the model to make noise-based decisions and thus lead to system overfitting.

Classification

Without feature selection

In this first experiment, the classification process was performed without feature selection, which means that all the attributes were considered and there was no pre-selection. The flow diagram of the global classification process is represented in Figure 3.4.

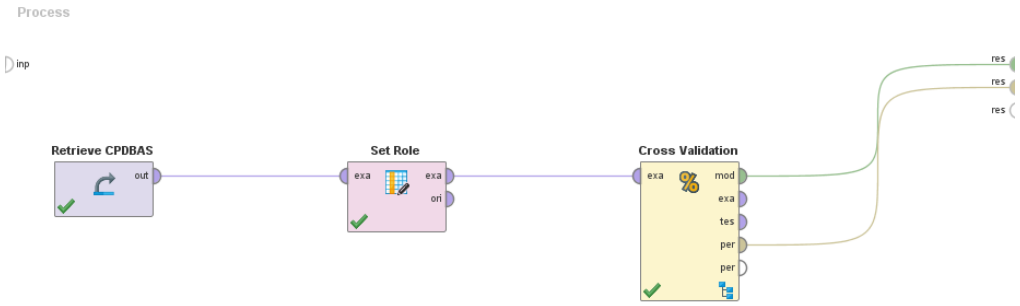


Figure 3.4: Classification process without feature selection.

The first step of the process was to open the data file through the "Retrieve" operator. This operator aims to read a file from the data repository.

The next operator used was the "Set Role" to change the target role of the '**Class**' (information extracted from DSSTox) attribute from "regular" to "label" because it was necessary to execute the classification algorithm. Thus, the classifier predicted this attribute. Figure 3.5 shows a sample of the output of this operator. When the attribute '**Class**' = active - means that the molecule has toxic characteristics, when the attribute '**Class**' = inactive - means that the molecule has no toxic properties.

Then we have used the "Cross Validation" operator. This operator automatically subdivides into the training and test procedures, as can be seen in Figure 3.6.

The training process contains the chosen classification operator. The testing process contains the "Apply Model" operator and the "Performance" operator.

"Apply Model" receives on the first port the model that contains all the information about the data with which it was trained and on the second port receives the test file.

The "Performance (Binominal classification)" operator was used to evaluate the model, in this case, was used to get the values of Accuracy, Precision, Recall and F-measure metrics.

ExampleSet (1230 examples, 1 special attribute, 506 regular attributes) Filter (1,230 / 1,230 examples): all

Row No.	Class	Name	nAcid	apol	naAromAtom	nAromBond	nAtom	nHeavyAtom	nH	nB	nC
1	active	48414658	0	24.841	0	0	16	10	6	0	4
2	active	48414657	0	16.181	0	0	12	6	6	0	3
3	active	48414643	0	41.306	12	12	35	20	15	0	15
4	active	48414642	0	22.515	6	6	20	9	11	0	8
5	active	48414639	0	20.748	6	6	18	8	10	0	8
6	active	48414633	0	109.224	24	24	90	50	40	0	39
7	active	48414632	0	18.463	0	0	17	8	9	0	6
8	active	48414627	0	6.077	0	0	6	3	3	0	2
9	active	48414626	0	7.700	0	0	6	3	3	0	2
10	active	48414624	0	8.570	0	0	6	3	3	0	2
11	active	48414623	0	12.645	0	0	12	6	6	0	4
12	active	48414619	0	12.652	0	0	13	6	7	0	3
13	active	48414616	0	13.511	0	0	12	8	4	0	4
14	active	48414609	0	31.755	13	15	26	15	11	0	12
15	active	48414608	0	34.848	13	15	29	16	13	0	13
16	active	48414604	0	45.760	0	0	35	20	15	0	9

Figure 3.5: Sample of the output of the "Set Role" operator for CPDBAS data set.

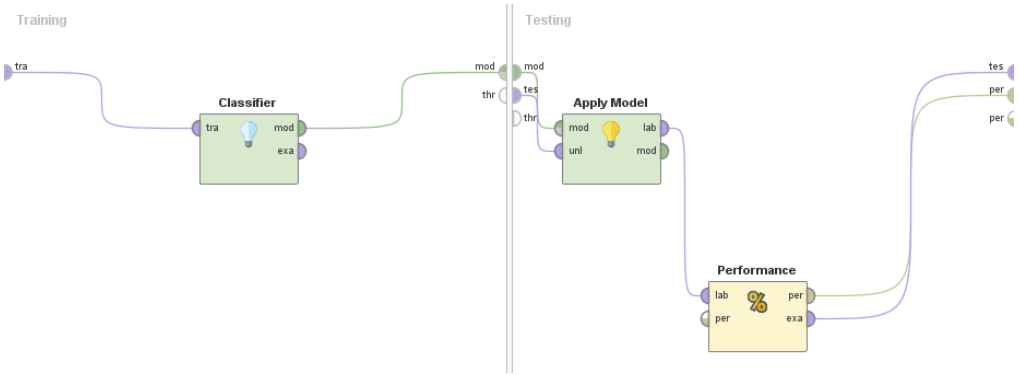


Figure 3.6: Sub-process of "Cross Validation" operator.

With feature selection

The first experiment was repeated with feature selection, which means, with a pre-selection of the best attributes. The flow diagram of the classification process with feature selection is represented in Figure 3.7.

As in previous experience, the first step was to open the data file through the "Retrieve" operator. The next operator used was the "Set Role". Once again this was used to change the target role of the '**Class**' attribute from "regular" to "label".

Next, the "Normalize" operator was applied to all attributes of the database. This step causes all attributes to be transformed so that they all vary between the same values. In this way, attributes with large values are prevented from having a greater weight in the prediction of each data example.

Afterwards, feature selection was made and two operators were used.

"Weight by Correlation" operator generates a weight for each attribute of the database, and this weight corresponds to the correlation between each input attribute and the label. The greater the

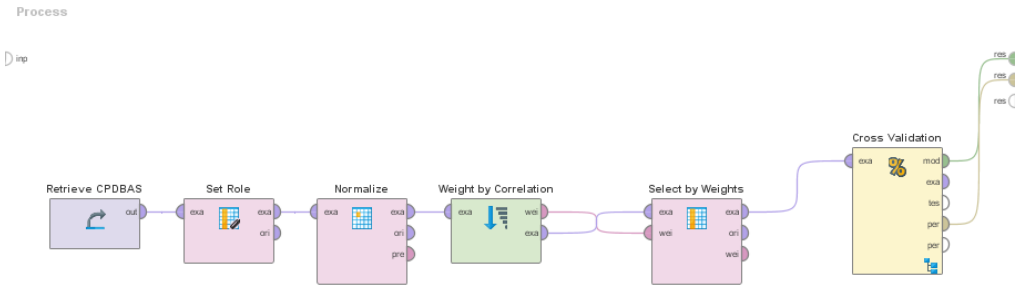
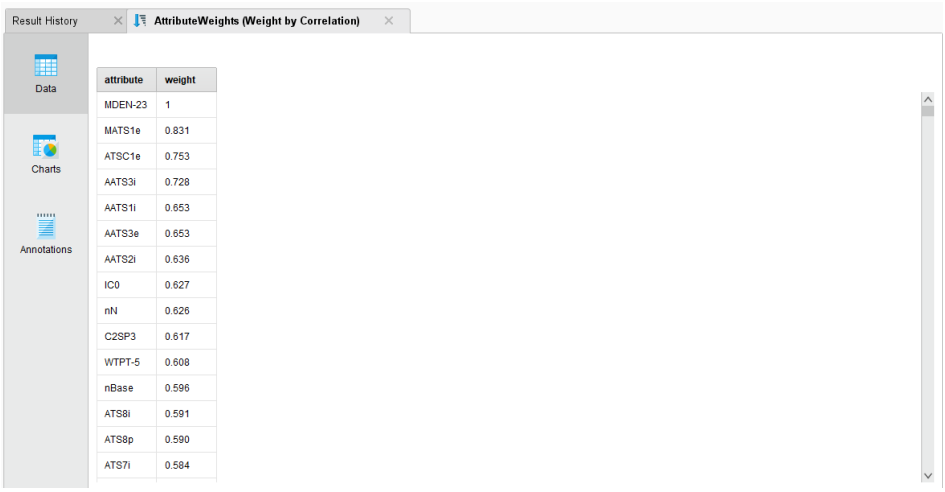


Figure 3.7: Classification process with feature selection.

weight of an attribute, the more relevance it has, and therefore it is taken in greater consideration to solve the problem. We show in Figure 3.8 a sample of the output of the "Weight by Correlation" operator.



attribute	weight
MDEN-23	1
MATS1e	0.831
ATSC1e	0.753
AATS3i	0.728
AATS1i	0.653
AATS3e	0.653
AATS2i	0.636
IC0	0.627
nN	0.626
C2SP3	0.617
WTPT-5	0.608
nBase	0.596
ATS8i	0.591
ATS8p	0.590
ATS7i	0.584

Figure 3.8: Sample of the output of the "Weight by Correlation" operator for CPDBAS data set.

"Select by Weights" operator selects only the "k" attributes with the highest weight (in this case, k = 10).

According to this operator, the top ten attributes of CPDBAS data set (in descending order of relevance) are:

MDEN-23- molecular distance edge between all secondary and tertiary nitrogens;

MATS1e- moran autocorrelation - lag 1 / weighted by Sanderson electronegativities;

ATSC1e- centered Broto-Moreau autocorrelation - lag 1 / weighted by Sanderson electronegativities;

AATS3i- average Broto-Moreau autocorrelation - lag 3 / weighted by first ionization potential;

AATS1i- average Broto-Moreau autocorrelation - lag 1 / weighted by first ionization potential;

AATS3e- average Broto-Moreau autocorrelation - lag 3 / weighted by Sanderson electronegativities;

AATS2i- average Broto-Moreau autocorrelation - lag 2 / weighted by first ionization potential;

IC0- information content index (neighborhood symmetry of 0-order);

nN- number of nitrogen atoms;

C2SP3- singly bound carbon bound to two other carbons[2].

It is important to note that these ten attributes were selected (by the operator) as the best for the four classification algorithms used in the experiment.

Then we have used the "Cross Validation" operator. As mentioned before, this operator is automatically subdivided into the training and test procedures.

The training process contains the chosen classification operator. The testing process contains the "Apply Model" operator and the "Performance (Binominal classification)" operator. This last one was used to get the values of Accuracy, Precision, Recall and F-measure.

The other toxicity experiment was similar to the previous one but for the EPAFHM data set.

The flowchart is the same, and the top ten attributes (in descending order of relevance) chosen by the "Select by Weights" operator are:

LipinskiFailures- number failures of the Lipinski's Rule Of 5;

ATS8v- broto-Moreau autocorrelation - lag 8 / weighted by van der Waals volumes;

WPATH- broto-Moreau autocorrelation - lag 8 / weighted by van der Waals volumes;

ATS8p- broto-Moreau autocorrelation - lag 8 / weighted by polarizabilities;

EE_D- estrada-like index from topological distance matrix;

SpMax_D- leading eigenvalue from topological distance matrix;

SpAD_D- spectral absolute deviation from topological distance matrix;

ATS7v- broto-Moreau autocorrelation - lag 7 / weighted by van der Waals volumes;

SpDiam_D- spectral diameter from topological distance matrix;

ECCEN- a topological descriptor combining distance and adjacency information[2] .

In the same way, it was also obtained the top ten for EPAFHM data filtered with tanimoto coefficient:

GGI8- topological charge index of order 8;

[2] www.scbdd.com/padel_desc/descriptors/

ATS8v- broto-Moreau autocorrelation - lag 8 / weighted by van der Waals volumes;

ZMIC1- Z-modified information content index (neighborhood symmetry of 1-order);

ATS8m- broto-Moreau autocorrelation - lag 8 / weighted by mass;

WPATH- weiner path number;

GGI9- topological charge index of order 9;

GGI7- topological charge index of order 7;

ATS8p- broto-Moreau autocorrelation - lag 8 / weighted by polarizabilities;

LipinskiFailures- number failures of the Lipinski's Rule Of 5;

EE_D- estrada-like index from topological distance matrix[3].

Once again, it is important to note that these ten attributes were selected (by the operator) as the best for the four classification algorithms used in the experiment.

3.3.2 Blood-Brain Barrier penetration Experiments

Experiments with BBB penetration data were performed using two Data Mining tasks, regression and classification (the difference between these two has already been explained in section 2.3).

The global work plan used for regression task can be seen in Table 3.2:

Table 3.2: Methodology for regression experiments of BBB penetration data set.

BBBDataSets = {Brain Blood Barrier}
DMalgorithms = {Linear Regression, SVM, k-NN}

Forall DS in **BBBDataSets**
  Enrich DS with Molecular Descriptors
  Forall Alg in **DMalgorithms**
    build and evaluate models
  EndFor

EndFor

The work plan used for the classification is similar to the regression, but with classification algorithms. This work plan is represented in Table 3.3:

Data set

Cheminformatics is a web repository that contains links to several data sets (already presented in section 2.2). The Blood-Brain-Barrier penetration data[4] used were the same as Liu et al. [72]

[3] www.scbdd.com/padel_desc/descriptors/
[4] www.cheminformatics.org/

Table 3.3: Methodology for classification experiments of BBB penetration data set.

```
BBBDataSets = {Blood-Brain Barrier}
DMalgorithms = {SVM & k-NN & Decision trees & Random forest}

Forall DS in BBBDataSets
  Enrich DS with Molecular Descriptors
  Forall Alg in DMalgorithms
    build and evaluate models
  EndFor

EndFor
```

used for their penetration studies. That studies were performed for regression task. However, the methodologies of evaluation of the models were different, and therefore it was not possible to compare their results with the results obtained in this experience.

Data pre-processing

The original data set was in SDF format, so for the application of the regression task it was necessary to originate the SMILE format of molecules (using OpenBabel) and calculate molecular descriptors (again using the PaDEL Descriptor software).

This data set contained 57 compounds as training set and 13 as test set. After the calculation of molecular descriptors (1D+2D), the data set got 1444 attributes.

After removing the missing values, BBB penetration data set got 1356 attributes.

For the classification task, the pre-processing described earlier was the same.

The original BBB penetration data is a regression problem [72], which means that the class is numeric. To apply the classification task it was necessary to transform the class into binominal.

The histogram of the activity of the molecules of the BBB penetration data was used to know where to divide the two classes. The point where there was a more pronounced difference between the points on the left side and the points on the right side was chosen.

A complete division and a division with a gray zone in the middle was made. This gray area was left to sharpen the boundary between the two classes (molecules with high capacity to penetrate the BBB and molecules with low capacity to penetrate the BBB), and thus help the classifiers.

This histogram is represented in Figure 3.9.

After analyzing the obtained histogram, a division was made to separate the values of the molecules and then create the label attribute for classification.

For division without the gray area:

values of activity ≥ 0 - classified as high;

values of activity < 0 - classified as low.

For division with the gray area:

values of activity ≥ 0.5 - classified as high;

values of activity < 0 - classified as low.

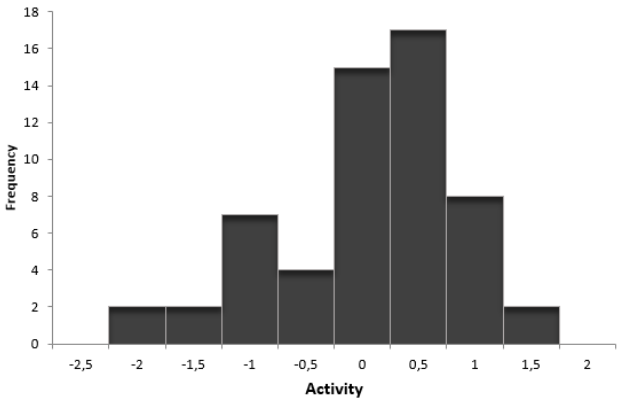


Figure 3.9: Representative histogram of the activity of molecules.

This gray zone comprises the values ≥ 0 and < 0.5.

When the attribute '**class**' = high - means that the molecule has high capacity to penetrate through BBB, when the attribute '**class**' = low - means that the molecule has low capacity to penetrate through BBB.

Figure 3.10 contains a sample of '**class**' attribute.

ExampleSet (57 examples, 0 special attributes, 1356 regular attributes) Filter (57 / 57 examples): all

Row No.	Name	class	nAcid	ALogP	ALogp2	AMR	apol	naAromAtom	nAromBond	nA
1	AUTOGEN_tr...	low	0	0.412	0.170	54.515	37.769	5	5	33
2	AUTOGEN_tr...	low	0	0.561	0.314	24.726	21.434	5	5	18
3	AUTOGEN_tr...	high	0	-0.520	0.271	43.378	60.734	18	18	53
4	AUTOGEN_tr...	high	0	-0.319	0.102	61.195	72.241	21	22	61
5	AUTOGEN_tr...	high	0	0.121	0.015	66.928	65.634	17	17	57
6	AUTOGEN_tr...	high	0	1.167	1.363	32.500	29.501	6	6	23
7	AUTOGEN_tr...	high	0	0.214	0.046	38.642	49.358	12	12	44
8	AUTOGEN_tr...	high	0	0.091	0.008	39.256	51.643	12	12	45
9	AUTOGEN_tr...	high	0	0.023	0.001	66.372	47.255	5	5	43
10	AUTOGEN_tr...	low	0	0.514	0.264	64.475	42.869	5	5	36
11	AUTOGEN_tr...	low	0	-0.953	0.907	21.814	30.442	11	12	27
12	AUTOGEN_tr...	high	0	1.532	2.348	58.375	43.239	6	6	35
13	AUTOGEN_tr...	high	0	1.333	1.778	41.019	41.742	11	11	33
14	AUTOGEN_tr...	high	0	0.445	0.198	32.420	39.359	11	11	33
15	AUTOGEN_tr...	high	0	0.901	0.812	37.021	55.680	17	17	46
16	AUTOGEN_tr...	high	0	0.121	0.015	18.993	31.568	11	12	25

Figure 3.10: Sample of the file corresponding to the attribute '**class**'.

Regression problem

The first experience with the BBB penetration data was performed for the regression. The flow diagram was constructed and is represented in the Figure 3.11.

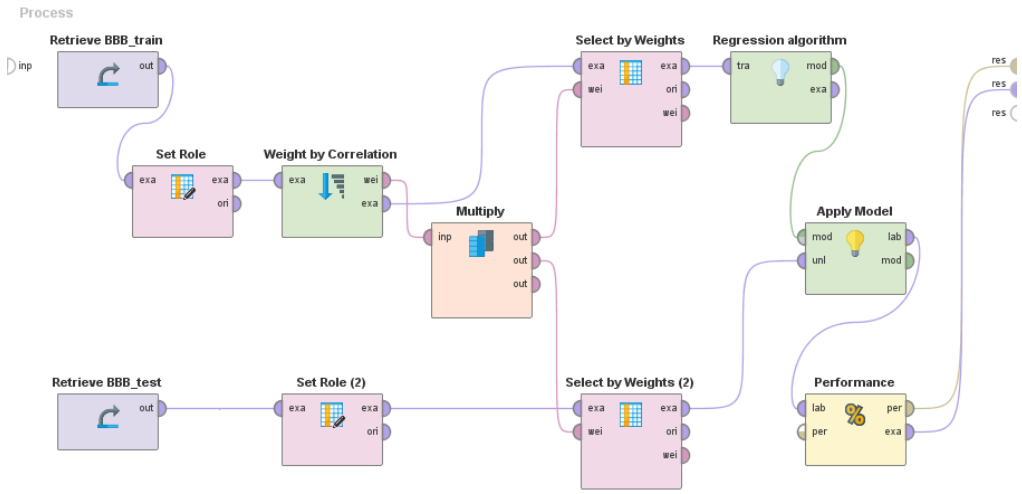


Figure 3.11: Linear Regression process.

The first step of the project was to open the data file through the "Retrieve" operator. This operator aims to read a file from the data repository. In this case, the database used had two files (test and training) and so it was necessary to read both.

The next operator used was the "Set Role" to change the target role of the '**logBBexpt**'[5] attribute from "regular" to "label" in order to execute successfully the regression algorithm. Thus, this regression algorithm predicted this attribute. Since there were two input data files, this operator was applied to both.

Feature selection was performed and the operators used were "Weight by Correlation" and "Select by Weights".

"Select by Weights" operator selects only the "k" attributes with the highest weight (in this case, k = 10).

According to this operator, the top ten attributes for BBB penetration data (in descending order of relevance) are:

nHBAcc- number of hydrogen bond acceptors (using CDK HBondAcceptorCountDescriptor algorithm);

TopoPSA- topological polar surface area;

nHBd- count of E-States for (strong) Hydrogen Bond donors;

nHBDon- number of hydrogen bond donors (using CDK HBondDonorCountDescriptor algorithm);

nHBAcc_Lipinski- number of hydrogen bond acceptors (using Lipinski's definition: any nitrogen; any oxygen);

nBondsD- number of double bonds;

[5]Corresponds to the logarithm of the ratio of the concentration of a drug in the brain and in the blood (this ratio is measured at equilibrium) and this value is an index of BBB permeability. logBBexpt means logBB experimental value.

nBondsD2- total number of double bonds (excluding bonds to aromatic bonds);

nHBint2- count of E-State descriptors of strength for potential Hydrogen Bonds of path length 2);

nN- number of nitrogen atoms;

SHBd- sum of E-States for (strong) hydrogen bond donors[6].

The "Multiply" operator was used to copy the input (weights) to multiple outputs (in this case for training and testing data).

Then the chosen regression algorithm was applied in order to do the numerical prediction.

Next, the "Apply Model" operator was used. This receives on the first port the training model that contains all the information about the data with which it was trained and on the second port receives the test file.

Finally, the "Performance (Regression)" operator was introduced. This was used to evaluate the model, in this case, to get the values of Root Mean Squared Error and Mean Absolute Error.

Classification problem

This classification experiment was performed for the two different data sets. A complete data set and another with the gray area (already explained in data pre-processing).

The flow diagram of the classification process was constructed and is represented in the Figure 3.12. This flowchart is similar to the regression one but with a classification algorithm and with the "Performance (Classification)" operator.

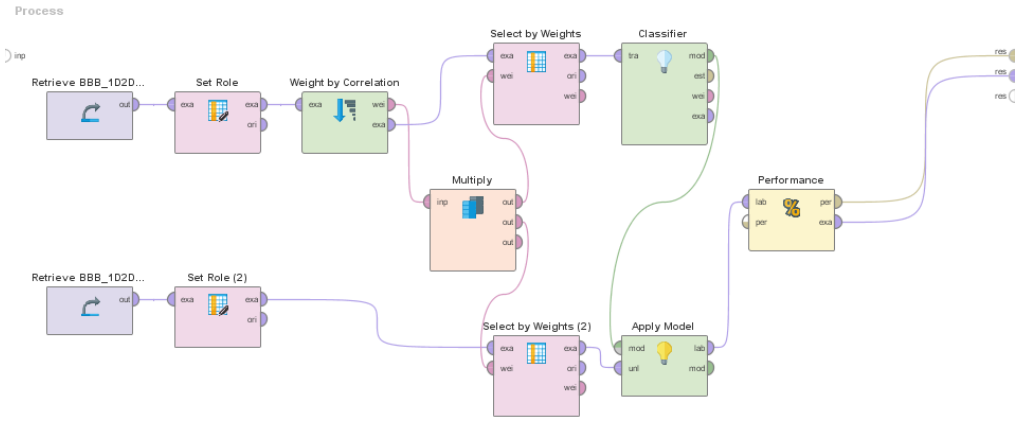


Figure 3.12: Classification process for BBB penetration data.

As in the classification for toxicity experiments, this flowchart was repeated for the four classification algorithms (SVM, k-NN, Decision Trees and Random Forest).

According to "Select by weights" operator, the top ten attributes for BBB penetration data without gray area (in descending order of relevance) are:

[6] www.scbdd.com/padel_desc/descriptors/

maxHBd- maximum E-States for (strong) Hydrogen Bond donors;

minHBd- minimum E-States for (strong) Hydrogen Bond donors;

MLFER_A- overall or summation solute hydrogen bond acidity;

SHBd- sum of E-States for (strong) hydrogen bond donors;

nHBd- count of E-States for (strong) Hydrogen Bond donors;

nHBDon- number of hydrogen bond donors (using CDK HBondDonorCountDescriptor algorithm);

TopoPSA- topological polar surface area;

nHBAcc_Lipinski- number of hydrogen bond acceptors (using Lipinski's definition: any nitrogen; any oxygen);

ETA_dEpsilon_D- a measure of contribution of hydrogen bond donor atoms;

nHBDon_Lipinski- number of hydrogen bond donors (using Lipinski's definition: Any OH or NH. Each available hydrogen atom is counted as one hydrogen bond donor)[7].

By the same way, according to "Select by weights" operator, the top ten attributes for BBB penetration data with gray area (in descending order of relevance) are:

ETA_Epsilon_2- a measure of electronegative atom count;

hmax- maximum H E-State;

AATS0e- average Broto-Moreau autocorrelation - lag 0 / weighted by Sanderson electronegativities;

ETA_dEpsilon_C- a measure of contribution of electronegativity;

ETA_dPsi_A- a measure of hydrogen bonding propensity of the molecules;

ETA_Psi_1- a measure of hydrogen bonding propensity of the molecules and/or polar surface area;

ETA_Epsilon_4- a measure of electronegative atom count;

Mse- mean atomic Sanderson electronegativities (scaled on carbon atom);

MAXDN- maximum negative intrinsic state difference in the molecule (related to the nucleophilicity of the molecule). Using deltaV = (Zv-maxBondedHydrogens) /(atomicNumber-Zv-1);

AATSC0e- average centered Broto-Moreau autocorrelation - lag 0 / weighted by Sanderson electronegativities[7].

The selection of attributes by this operator is important because it helps to understand which descriptors most significantly influence the activity of the molecule.

[7] www.scbdd.com/padel_desc/descriptors/

Chapter 4

Results and discussion

This chapter presents the global results of the experiments. Toxicity experiments were performed using a classification task and the results of the experiment are described in the first part of the chapter. The penetration experiments were performed with two tasks, regression and classification, and the results obtained are described in the second part of the chapter. A discussion / evaluation of these results is also made. The metrics used to evaluate the results were Accuracy, Recall, Precision and F-measure for classification, and Root Mean Squared Error and Mean Absolute Error for regression.

4.1 Toxicity Experiments

Toxicity experiments had as objective the use of classification algorithms to predict the behavior of the molecules for toxicity. For the accomplishment of this task, the classication algorithms used were:

- Support Vector Machines (tested with kernel Radial Basis Function (RBF) adjusted in the parameters);

- k-NN (the parameter of k-NN (k) was tested for the values of 1, 3 and 5. The most satisfactory parameter was k = 5 and so was used in the several experiments);

- Decision tree with CART;

- Random Forest with CART (the number of trees used was equal to 10).

Classification

Without feature selection

In this experiment no feature selection was performed and all attributes were considered. For a better understanding of the tables, it is important to note that the first column represents the evaluation metrics: Accuracy, Precision (active) - corresponds to the precision of the active class (molecules with toxic activity) -, Recall (active) - corresponds to the recall of the active class -, Precision (inactive) - corresponds to the precision of the inactive class (molecules without toxicity

activity) -, Recall (inactive) - corresponds to the recall of the inactive class - and F-measure. The other four columns represent the four classification algorithms used.

Table 4.1 shows the results obtained for CDPBAS data set without feature selection, for all the evaluation metrics and the four different classification algorithms. This experiment was performed using 10-fold Cross Validation (k = 10).

Table 4.1: Results of the classification experiment (without feature selection) for CPDBAS data set.

Evaluation metrics	SVM	k-NN	Decision Tree	Random Forest
Accuracy	52.28% (+/- 1.41%)	**56.99% (+/- 4.82%)**	50.98% (+/- 1.86%)	51.95% (+/- 2.94%)
Precision (active)	48.57%	54.77%	47.81%	46.03%
Recall (active)	2.90%	55.80%	31.74%	4.95%
Precision (inactive)	52.38%	59.08%	52.44%	52.27%
Recall (inactive)	97.20%	58.07%	68.48%	94.72%
F-measure	68.08% (+/- 0.99%)	58.65% (+/- 3.72%)	52.70% (+/- 21.05%)	67.35% (+/- 2.25%)

Taking advantage of the Decision Tree algorithm not be an algorithm of the black box type, a survey was made regarding the molecular descriptors chosen as root of the decision tree.

Inspection to the generated decision tree for the original CPDBAS data set, revealed that the best descriptor (tree root) was MDEN-23. This descriptor means the molecular distance edge between all secondary and tertiary nitrogens[1].

After analyzing the values obtained for all the metrics, we verified that the classification algorithm **k-NN** was the one that obtained more satisfactory prediction results, with an accuracy of 56.99% for CPDBAS data set.

With feature selection

Table 4.2 shows the results obtained for CPDBAS data set with feature selection, for all the evaluation metrics and the four different classification algorithms. This experiment was performed using 10-fold Cross Validation (k = 10).

Table 4.2: Results of the classification experiment (with feature selection) for CPDBAS data set.

Evaluation metrics	SVM	k-NN	Decision Tree	Random Forest
Accuracy	**60.81% (+/- 2.93%)**	58.54% (+/- 4.97%)	52.28% (+/- 0.37%)	53.41% (+/- 2.49%)
Precision (active)	67.45%	55.72%	44.44%	62.26%
Recall (active)	34.30%	63.14%	0.68%	5.63%
Precision (inactive)	58.69%	61.84%	52.33%	53.02%
Recall (inactive)	84.94%	54.35%	99.22%	96.89%
F-measure	69.33% (+/- 3.04%)	57.70% (+/- 6.23%)	68.52% (+/- 0.34%)	68.55% (+/- 0.77%)

Observing the generated decision tree for the CPDBAS data set, we found that the best descriptor (tree root) was AATS3i. This descriptor means average Broto-Moreau autocorrelation - lag 3 / weighted by first ionization potential[1].

[1] www.scbdd.com/padel_desc/descriptors/

After analyzing the values obtained for all the metrics, we verified that the classification algorithm **SVM** was the one that obtained more satisfactory prediction results, with an accuracy of 60.81% for CPDBAS data set.

Comparing the results obtained by the two experiments (without and with feature selection), we note that accuracy values are higher when feature selection is performed, which was expected, since a pre-selection of the best attributes was made. Figure 4.1 shows a comparison of this two experiments.

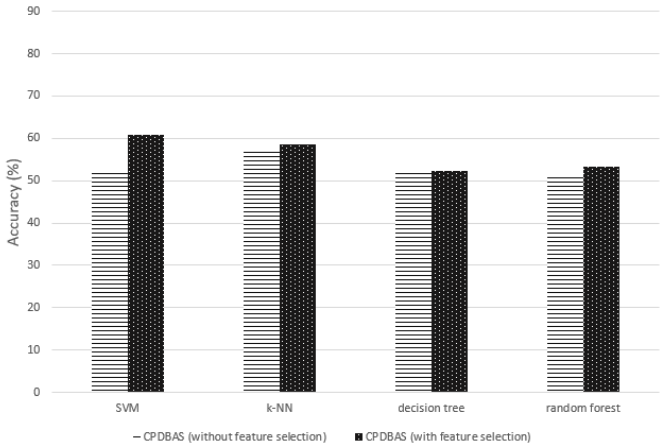


Figure 4.1: Results of the accuracy for CPDBAS data set (without and with feature selection) and the four classification algorithms under study.

Through the graph it is possible to verify that the data that obtained the best accuracy was CPDBAS with feature selection when performed with **SVM** algorithm. The least favorable result was obtained for CPDBAS without feature selection data set when performed with Random Forest algorithm.

As previously mentioned, an overall analysis of the graph allows to conclude that the four algorithms work better when the feature selection is performed.

Table 4.3 presents the global accuracy averages (without feature selection and with feature selection) for the four classification algorithms.

Table 4.3: Results of the accuracy averages for each of the algorithms used.

Algorithm	Accuracy average
SVM	56.55%
k-NN	**57.77%**
Decision Tree	52.12%
Random Forest	52.20%

Briefly, the algorithm **k-NN** was the one that achieved the best accuracy (57.77%) followed by SVM, with an accuracy of 56.55%.

As previously mentioned, the CPDBAS data set was filtered with tanimoto coefficient. However, there were no changes to the original file, which proved that the molecules are sufficiently differentiated and therefore the results of the experiment were not presented.

To compare the original and tanimoto-filtered data, it was used EPAFHM data set. After filtering with tanimoto coefficient, this data set was composed of 396 compounds (instead of the original 617).

Table 4.4 presents the results obtained by the **SVM** algorithm for EPAFHM data sets. Once again, this experiment was performed using 10-fold Cross Validation (k = 10).

Table 4.4: Results of the classification experiment (with feature selection) with **SVM** algorithm for the two EPAFHM data sets.

Evaluation metrics	EPAFHM	EPAFHM (tanimoto filtered)
Accuracy	94.31% (+/- 1.05%)	94.46% (+/- 1.85%)
Precision (active)	94.45%	94.42%
Recall (active)	99.83%	100.00%
Precision (inactive)	0%	100%
Recall (inactive)	0%	8.33%
F-measure	—	15.38%

Table 4.5 presents the results obtained by the **k-NN** algorithm. Once again, this experiment was performed using 10-fold Cross Validation (k = 10).

Table 4.5: Results of the classification experiment (with feature selection) with **k-NN** algorithm for the two EPAFHM data sets.

Evaluation metrics	EPAFHM	EPAFHM (tanimoto filtered)
Accuracy	94.96% (+/- 2.34%)	94.21% (+/- 3.35%)
Precision (active)	96.92%	95.56%
Recall (active)	97.76%	98.39%
Precision (inactive)	55.17%	53.85%
Recall (inactive)	47.06%	29.17%
F-measure	50.45%	37.84%

Table 4.6 shows the results obtained by the **Decision Tree** algorithm. Once again, this experiment was performed with 10-fold Cross Validation (k = 10).

Table 4.6: Results of the classification experiment (with feature selection) with **Decision Tree** algorithm for the two EPAFHM data sets.

Evaluation metrics	EPAFHM	EPAFHM (tanimoto filtered)
Accuracy	**95.45% (+/- 1.73%)**	95.22% (+/- 2.83%)
Precision (active)	96.15%	95.84%
Recall (active)	99.14%	99.19%
Precision (inactive)	68.75%	72.73%
Recall (inactive)	32.35%	33.33%
F-measure	44.00%	45.71%

Observing the generated decision tree for the original EPAFHM data set, we found that the best descriptor (tree root) was ATS8v. This descriptor means the broto-Moreau autocorrelation - lag 8 / weighted by van der Waals volumes[2].

Looking at the generated decision tree for EPAHFM tanimoto-filtered data set, we saw that the best descriptor (tree root) was GGI8. This descriptor means the topological charge index of order 8[2].

Table 4.7 shows the results obtained by the **Random Forest** algorithm. Once again, this experiment was performed using 10-fold Cross Validation (k = 10).

Table 4.7: Results of the classification experiment (with feature selection) with **Random Forest** algorithm for the two EPAFHM data sets.

Evaluation metrics	EPAFHM	EPAFHM (tanimoto filtered)
Accuracy	95.28% (+/- 1.13%)	94.70% (+/- 2.37%)
Precision (active)	95.12%	95.12%
Recall (active)	98.97%	99.46%
Precision (inactive)	64.71%	71.43%
Recall (inactive)	32.35%	20.83%
F-measure	43.14%	32.26%

After analyzing the previous tables for, we saw that the data set that had a higher accuracy value (95,45%) was the original EPAFHM when performed with **Decision Tree** algorithm. So, it is also possible to note that the data filtered with tanimoto coefficient had slightly lower accuracy values than the original data.

The graph represented in the Figure 4.2 shows a comparison of classification task for EPAFHM original and EPAFHM tanimoto-filtered .

Through the graph it is possible to verify that the data that obtained the best accuracy was EPAFHM original data set.

It is normal that the original data set presents better results than tanimoto-filtered data set, because the first one has more similar molecules. By removing similar molecules, the classification problem becomes more complicated.

The filtering with the tanimoto coefficient aims to obtain more realistic results.

[2] www.scbdd.com/padel_desc/descriptors/

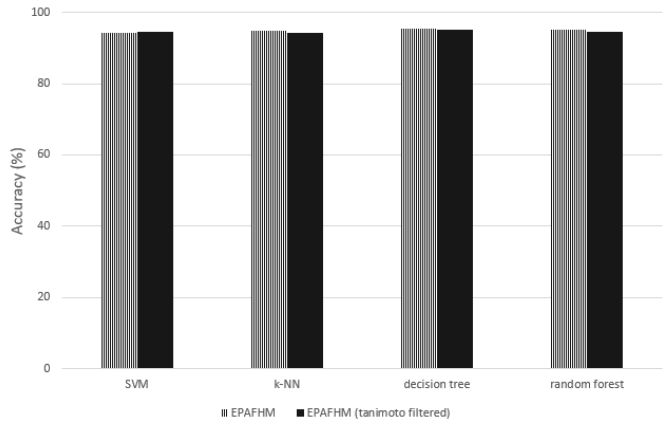


Figure 4.2: Results of the accuracy for the two EPAFHM data sets and the four classification algorithms.

4.2 Blood-Brain Barrier penetration Experiments

Regression problem

This experiment was carried out with the objective of predicting the behavior of the molecules (high penetration capacity or low penetration capacity) in the process of Rational Drug Design. For the accomplishment of this task, the algorithms used were:

- Linear Regression;

- Support Vector Machines (it was adjusted in the parameters for 'nu-SVR svm type' in order to be used for regression task);

- k-NN (the parameter of k-NN (k) was adjusted for k = 5).

The results of the regression experiment with **Linear regression** algorithm are shown in the Table 4.8.

Table 4.8: Results of the regression experiment with **Linear regression** algorithm for BBB penetration data set.

Evaluation metrics	BBB penetration
Root Mean Squared Error	0.779
Mean Absolute Error	0.671

The results of the regression experiment with **SVM** algorithm are shown in the Table 4.9.

Table 4.9: Results of the regression experiment with **SVM** algorithm for BBB penetration data set.

Evaluation metrics	BBB penetration
Root Mean Squared Error	0.610
Mean Absolute Error	0.453

The results of the regression experiment with **k-NN** algorithm are shown in the Table 4.10.

Table 4.10: Results of the regression experiment with **k-NN** algorithm for BBB penetration data set.

Evaluation metrics	BBB penetration
Root Mean Squared Error	0.649
Mean Absolute Error	0.468

After analyzing the obtained values, it is possible to verify that the best results for the two evaluation metrics are obtained when using **SVM** algorithm (RSME = 0.610 and MAE = 0.453).

Classification problem

As mentioned before, the regression task was converted into a classification task. Once again, for the accomplishment of this experiment, the classication algorithms used were:

- Support Vector Machines (tested with kernel Radial Basis Function (RBF) adjusted in the parameters);

- k-NN (the parameter of k-NN (k) was tested for the values of 1, 3 and 5. The most satisfactory parameter was k = 5 and so was used in the several experiments);

- Decision tree with CART;

- Random Forest with CART (the number of trees used was equal to 10).

Without gray area

For a better understanding of the table, it is important to note that the first column represents the evaluation metrics: Accuracy, Precision (low) - corresponds to the precision of the low class (molecules with low capacity to penetrate BBB) -, Recall (low) - corresponds to the recall of the low class -, Precision (high) - corresponds to the precision of the high class (molecules with high capacity to penetrate BBB) -, Recall (high) - corresponds to the recall of the high class - and F-measure. The other four columns represent the four classification algorithms used.

Table 4.11 shows the results obtained for BBB penetration data set (without gray area), for all the evaluation metrics and the four classification algorithms.

Table 4.11: Results of the classification experiment (without gray area) for Blood-Brain Barrier penetration data set.

Evaluation metrics	SVM	k-NN	Decision Tree	Random Forest
Accuracy	**69.23%**	**69.23%**	61.54%	61.54%
Precision (low)	80.00%	100.00%	77.78%	77.78%
Recall (low)	80.00%	60.00%	70.00%	70.00%
Precision (high)	33.33%	42.86%	25.00%	25.00%
Recall (high)	33.33%	100.00%	33.33%	33.33%

Looking at the generated decision tree for the BBB (without gray area) data set, we verified that the best descriptor (tree root) was maxHBD. This descriptor means the maximum E-States for (strong) Hydrogen Bond donors [3].

By observing the table, it is possible to verify that the algorithms that obtained the best performance were **k-NN** and **SVM** with an accuracy value of 69.23% for BBB penetration data set without gray area.

With gray area

Table 4.12 shows the results obtained for BBB penetration data set (with gray area), for all the evaluation metrics and the four classification algorithms.

Table 4.12: Results of the classification experiment (with gray area) for Blood-Brain Barrier penetration data set.

Evaluation metrics	SVM	k-NN	Decision Tree	Random Forest
Accuracy	**90.91%**	72.73%	81.82%	81.82%
Precision (low)	90.91%	100.00%	100.00%	100.00%
Recall (low)	100.00%	70.00%	80.00%	80.00%
Precision (high)	0%	25.00%	33.33%	33.33%
Recall (high)	0%	100.00%	100.00%	100.00%

After observing the generated decision tree for the BBB (without gray area) data set, we saw that the best descriptor (tree root) was AATS0e. This descriptor means the average Broto-Moreau autocorrelation - lag 0 / weighted by Sanderson electronegativities[4].

By observing the table it is possible to verify that the algorithm with the best performance was **SVM** with an accuracy value of 90.91% for BBB penetration data set with gray area.

Comparing the two data groups, it is possible to verify that the BBB penetration data with the gray area (accuracy = 90.91%) produced more promising results than the data without the gray area (accuracy = 69.23%).

Figure 4.3 shows a comparison of the accuracy for the two data sets considered and the four algorithms used.

Once again, it can be seen in the bar graph that for all algorithms, BBB penetration data set with gray area was the one with the best accuracy (90.91%) for **SVM** algorithm.

As the best result was obtained for the BBB penetration data set with gray area for the **SVM** algorithm, the experiment was repeated using the "Cross Validation" operator to evaluate the model. This experiment was performed using 10-fold Cross Validation (k = 10).

The flow diagram applied was the same as for toxicity experiments and the results are presented in Table 4.13.

Table 4.13: Results of the classification experiment (with gray area) with **SVM** algorithm.

[3] www.scbdd.com/padel_desc/descriptors/

[4] www.scbdd.com/padel_desc/descriptors/

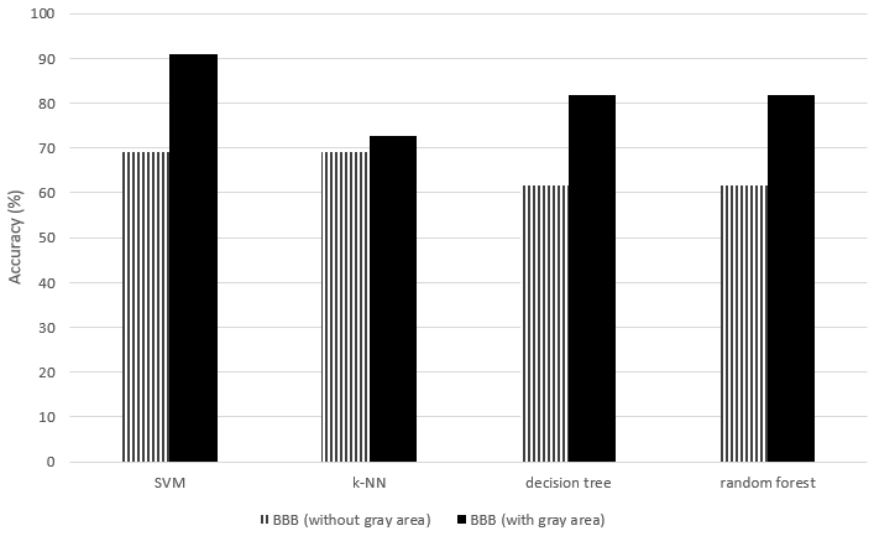


Figure 4.3: Accuracy for the four algorithms under study.

Evaluation metrics	BBB penetration
Accuracy	96.00% (+/- 8.00%)
Precision (low)	95.12%
Recall (low)	100.00%
Precision (high)	100.00%
Recall (high)	81.82%
F-measure	90%

By observing the table it is possible to verify that the accuracy result (accuracy = 96.00%) was superior to the result obtained when pre-dividing the training and test set (accuracy = 90.91%).

Chapter 5

Conclusions and Future work

In this chapter the conclusions of the work are presented and the results of the research are commented, as well as a proposal for future work.

5.1 Satisfaction of results

This work was motivated by the increasing appearance of diseases and the need to develop drugs (in the shortest possible time) to combat them. Research and development of new drugs is constantly evolving due to the effort of the pharmaceutical industry and to scientific development made at research institutions to bring better products to solve health problems.

A major hurdle in the development of a new drug is time (about 5-12 years) and the very high costs. When the process refers to the design of drugs for the Central Nervous System (CNS) can take even longer (up to 16 years). Informatics, and more specifically Data Mining technology may help to mitigate those hurdles. Based on historical information on pre-clinical trials results, Data Mining can build predictive models to estimate important features of the developed drugs, saving money and time by reducing / avoiding those pre-clinical tests.

The main objective of this dissertation was to asses how useful Data Mining methods and algorithms can be for the reduction of the Rational Drug Design process (more properly the ADMET tests phase). To meet this goal, it was first necessary to conduct a study of the various concepts associated with RDD and DM tools and algorithms that would be very useful for predicting drugs efficacy.

In order to answer the two research questions stated in section 1.2, two case studies were carried out. The first case study involved the realization of a classification experiment using two data sets downloaded from DSSTox database.

The first experiment corresponded to toxicity experiments. We have used four classification algorithms: SVM, k-NN, Decision Tree and Random Forest. The best results were obtained for the k-NN algorithm (accuracy = 57.77%). However, this value was not very promising. This experience was performed without feature selection, and then with feature selection (pre-selection

of the attributes). With a comparative analysis of the results obtained it was possibly to verify that the use of feature selection improved the performance of the models.

The second experiment used Blood-Brain Barrier penetration data. These experiments were performed for the regression and classification tasks. For regression it were used three algorithms: Linear regression, SVM and k-NN algorithms. The best results (lowest errors) were obtained for the SVM algorithm (RMSE = 0.610 and MAE = 0.453). Classification was also performed using the same four algorithms of the first experiment. We have considered two data sets, one complete (without gray area) and another considering a gray area between the two classes (high and low). Analyzing the results obtained, the best results were obtained for the data set with gray area for SVM algorithm (accuracy = 90.91%).

In response to the research question H1 (stated in section 1.2), the results suggest that Data Mining tool can be really useful to improve the drug design process. However, there are still many gaps that need to be addressed and improved.

Regarding the research question H2, it was not fully answered since the only additional information we used were the 1D and 2D molecular descriptors. It would be interesting to add information from 3D molecular descriptors, but it was not possible because the software could not calculate these descriptors.

This dissertation revealed the interest of using Data Mining tools and methods to aid in the Rational Drug Design process and the results obtained are encouraging for the continuation of this line of research.

5.2 Future work

A proposal to improve the model would be the inclusion of ontologies (described in section 2.4) as data enrichment. Thus, the inclusion of ontologies would be useful to verify if the addition of chemical information improves the performance of Data Mining algorithms by helping the generalization process.

Another interesting approach would be to compare the prediction results for different operators of feature selection (pre-selection of attributes) in order to verify if it would influence the performance of the model.

References

[1] Priscilla Regina Nasciutti. Desenvolvimento de novos fármacos, 2012. Universidade federal de Goiás, Escola de veterinária e zootecnia.

[2] N. Joan Abbott. Astrocyte–endothelial interactions and blood–brain barrier permeability. *Journal of Anatomy*, 200(6):629–638, 2002. URL: http://dx.doi.org/10.1046/j.1469-7580.2002.00064.x, doi:10.1046/j.1469-7580.2002.00064.x.

[3] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. Advances in knowledge discovery and data mining. chapter From Data Mining to Knowledge Discovery: An Overview, pages 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996.

[4] M. Hossin and M.N Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining Knowledge Management Process*, 5, March 2015.

[5] Joana Araújo Gonçalves. O circuito do medicamento: Da molécula à farmácia, 2011. Universidade Fernando Pessoa.

[6] Paulina Mata. Design racional de fármacos. *Química*, 1996.

[7] Adriano D. Andricopulo e Glaucius Oliva Rafael V. C. Guido. Planejamento de fármacos, biotecnologia e química medicinal: aplicações em doenças infecciosas. *Estudos avançados*, 24, 2010.

[8] Mohammad S. Alavijeh, Mansoor Chishty, M. Zeeshan Qaiser, and Alan M. Palmer. Drug metabolism and pharmacokinetics, the blood-brain barrier, and central nervous system drug discovery. *NeuroRx*, pages 554–571, October 2005.

[9] Timothy S. Carpenter, Daniel A. Kirshner, Edmond Y. Lau, Sergio E. Wong, Jerome P. Nilmeier, and Felice C. Lightstone. A method to predict blood-brain barrier permeability of drug-like compounds using molecular dynamics simulations. *Biophysical Journal*, 107:630–641, Aug 2014.

[10] David J. Begley. The blood-brain barrier: Principles for targeting peptides and drugs to the central nervous system. *Journal of Pharmacy and Pharmacology*, 48(2):136–146, 1996.

[11] Hugo Rojas, Cristiane Ritter, and Felipe Dal Pizzol. Mecanismos de disfunção da barreira hematoencefálica no paciente criticamente enfermo: ênfase no papel das metaloproteinases de matriz. *Rev Bras Ter Intensiva*, 23:222–227, 2011.

[12] N. Joan Abbott, Adjanie A.K. Patabendige, Diana E.M. Dolman, Siti R. Yusof, and David J. Begley. Structure and function of the blood–brain barrier. *Neurobiology of Disease*, 37:13–25, 2010.

[13] Li H, Yap CW, Ung CY, Xue Y, Cao ZW, and Chen YZ. Effect of selection of molecular descriptors on the prediction of bloodbrain barrier penetrating and nonpenetrating agents by statistical learning methods. *Journal of Chemical Information and Modeling*, 45(5):1376–1384, 2005. doi:10.1021/ci050135u.

[14] William M. Pardridge. The blood-brain barrier: Bottleneck in brain drug development. *NeuroRx*, 2:3–14, Jan 2005.

[15] William M. Pardridge. Blood-brain barrier drug targeting: The future of brain drug development. *Molecular interventions*, 3, 2003.

[16] Ikumi Tamai and Akira Tsuji. Transporter-mediated permeation of drugs across the blood–brain barrier. *Journal of Pharmaceutical Sciences*, 89:1371–1388, Nov 2000.

[17] Mohammad S. Alavijeh, Mansoor Chishty, M. Zeeshan Qaiser, and Alan M. Palmer. Drug metabolism and pharmacokinetics, the blood-brain barrier, and central nervous system drug discovery. *NeuroRX*, 2(4):554 – 571, 2005. Drug Discovery for Disorders of the Central Nervous System. URL: http://www.sciencedirect.com/science/article/pii/S1545534306701031, doi:https://doi.org/10.1602/neurorx.2.4.554.

[18] Andrew R.Leach and Valerie J.Gillet. *An Introduction to Chemoinformatics*. Springer, 2007.

[19] Berenice da Silva Junkes. Índice semi-empírico topológico: Desenvolvimento e aplicação de um novo descritor molecular em estudos de correlação quantitativa estrutura-propriedade (qspr). Technical report, Universidade Federal de Santa Catarina, Mai 2003.

[20] Mati Karelson and Victor S. Lobanov. Quantum-chemical descriptors in QSAR/QSPR studies. *American Chemical Society*, 96:1027–1043, Feb 1996.

[21] Roberto Todeschini and Viviana Consonni. *Handbook of Molecular Descriptors*, volume 11. WILEY-VCH, 2000.

[22] Sara Miguel Dinis Mamede da Cruz. Desenvolvimento de uma abordagem computacional para a descoberta de compostos-líderes para fármacos anticancerígenos, Set 2016. Universidade Nova de Lisboa, Faculdade de ciências e tecnologia.

[23] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open babel: An open chemical toolbox. *Journal of Cheminformatics*, Oct 2011.

[24] Z.R. Li, L.Y. Han, Y. Xue, C.W. Yap, H. Li, L. Jiang, and Y.Z. Chen. Model - molecular descriptor lab: A web-based server for computing structural and physicochemical features of compounds. *Biotechnology and Bioengineering*, 97:389 – 396, Jun 2007.

[25] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 4:33–38, Feb 1996.

[26] S. Sumathi and S.N. Sivanandam. *Introduction to Data Mining and its Applications*. Springer, 2006.

[27] Noemi Dreyer Galvão and Heimar de Fátima Marin. Data mining: a literature review. *Acta Paul Enferm*, 22:686–690, 2009.

[28] Gercely da Silva e Silva. Estudo de técnicas e utilização de mineração de dados em uma base de dados da saúde pública, 2003. Universidade Luterana do Brasil.

[29] Tauller Augusto de Araújo Matos. Uma visão geral das principais tarefas de mineração de dados, 2012. Fundação Educacional Dom André Arcoverde.

[30] Cássio Oliveira Camilo e João Carlos da Silva. Mineração de dados: Conceitos, tarefas, métodos e ferramentas, Ago 2009. Instituto de Informática, Universidade Federal de Goiás.

[31] S.M.Tondare S.D.Gheware, A.S.Kejkar. Data mining: Task, tools, techniques and applications. *International Journal of Advanced Research in Computer and Communication Engineering*, 3:8095–8098, Oct 2014.

[32] Colin Shearer. The crisp-dm model: The new blueprint for data mining. *Journal of Data Warehousing*, 5:13–22, 2000.

[33] Dr. Matthew North. *Data Mining for the Masses*. Global text, 2012.

[34] Yongjian Fu. Data mining: Tasks, techniques and applications, 1997. University of Missouri.

[35] RapidMiner. *RapidMiner 7. Operator Reference Manual*. 2016.

[36] Diana Colombo Pelegrin, Diego Paz Casagrande, Merisandra Côrtes de Mattos, Priscyla Waleska Targino de Azevedo Simões, Rafael Charnovscki, and Jane Bettiol. As tarefas de associação e de classificação na shell de data mining orion. 2013.

[37] Lissandra Luvizao Lazzarotto, Alcione de Paiva Oliveira, and Joelsio Jose Lazzarotto. Aspectos teóricos do data mining e aplicação das redes neurais em previsões de preços agropecuários. 2006.

[38] Yan yan Song and Ying LU. Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, 27:130–135, 2015.

[39] Aritz Pérez, Pedro Larrañaga, and Iñaki Inza. Bayesian classifiers based on kernel density estimation: Flexible classifiers. *International Journal of Approximate Reasoning*, 50:342–363, 2009.

[40] Vasile Paul Bresfelean, Mihaela Bresfelean, and Ramona Lacurezeanu. *Data Mining Tasks in a Student-Oriented DSS*, pages 321–328. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. doi:10.1007/978-3-642-25908-1_41.

[41] V.Sathiya Suntharam and Dr.Sai Satya Narayana Reddy. Data mining tasks performed by temporal sequential pattern. *International Journal of Research and Computational Technology*, 2, Jun 2012.

[42] Swati Gupta. A regression modeling technique on data mining. *International Journal of Computer Applications*, 116, Apr 2015.

[43] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10–18, 2009. URL: http://www.sigkdd.org/explorations/issues/11-1-2009-07/p2V11n1.pdf.

[44] Marcelino Pereira dos Santos Silva. Mineração de dados - conceitos, aplicações e experimentos com weka, 2004. Universidade do Estado do Rio Grande do Norte.

[45] Ian H. Witten and Eibe Frank. Data mining practical machine learning tools and techniques. pages 365–368, 2005.

[46] Tânia Gomes. Ferramentas open source de data mining, Dec 2014. Instituto Superior de Engenharia de Coimbra.

[47] Sérgio Francisco dos Santos Morais. Sistemas de recomendação em rapid miner: um caso de estudo, 2012. Faculdade de Economia da Universidade do Porto.

[48] Cagatay Catal. Performance evaluation metrics for software fault prediction studies. *Acta Polytechnica Hungarica*, 9, 2012.

[49] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, pages 427–437, May 2009.

[50] Cristina Oprea. Performance evaluation of the data mining classification methods. *Information society and sustainable development*, pages 249–253, 2014.

[51] Oliviero Carugo. Detailed estimation of bioinformatics prediction reliability through the fragmented prediction performance plots. *BMC Bioinformatics*, 8:380 – 380, 2007.

[52] P. Baldi, Søren Brunak, Y. Chauvin, and Henrik Nielsen. Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics*, 16(5):412–424, 2000. doi:10.1093/bioinformatics/16.5.412.

[53] Paola Gramatica and Alessandro Sangion. A historical excursus on the statistical validation parameters for qsar models: a clarification concerning metrics and terminology. *Journal of Chemical Information and Modeling*, May 2016.

[54] Isabel Durán Munoz and María Rosario Bautista Zambrana. Applying ontologies to terminology: Advantages and disadvantages. *Journal of Language and Communication in Business*, pages 64–76, December 2013.

[55] Edison Andrade Martins Morais and Ana Paula L. Ambrósio. Ontologias: conceitos, usos, tipos, metodologias, ferramentas e linguagens. 2007.

[56] B. Chandrasekaran, John R. Josephson, and V. Richar Benjamins. What are ontologies, and why do we need them? *Journal of Biomedical Semantics*, pages 20–26, February 1999.

[57] Willem Nico Borst. Construction of engineering ontologies for knowledge sharing and reuse, 1997. University of Twente, Enschede.

[58] E. A. M. Morais and A. P. L. Ambrósio. Ontologias: conceitos, usos, tipos, metodologias, ferramentas e linguagens, Dec 2007. Instituto de Informática, Universidade Federal de Goiás.

[59] Frederico Luiz Gonçalves de Freitas. Ontologias e a web semântica, 2004. Universidade Católica de Santos.

[60] Anarosa Alves Franco Brandão e Carlos José Pereira de Lucena. Uma introdução à engenharia de ontologias no contexto da web semântica, Nov 2002. Pontifícia Universidade Católica do Rio de Janeiro.

[61] Paula de Matos, A Dekker, M Ennis, Janna Hastings, K Haug, S Turner, and Christoph Steinbeck. Chebi: a chemistry ontology and database. *Journal of Cheminformatics*, May 2010.

[62] Janna Hastings, Despoina Magka, Colin Batchelor, Lian Duan, Robert Stevens, Marcus Ennis, and Christoph Steinbeck. Structure-based classification and ontology in chemistry. *Journal of Cheminformatics*, 2012.

[63] Janna Hastings, Kirill Degtyarenko, Paula de Matos, Marcus Ennis, and Christoph Steinbeck. The chebi ontology: an ontology for chemistry within a biological context. 2010.

[64] V. Maniraj and Dr.R. Sivakumar. Ontology languages – a review. *International Journal of Computer Theory and Engineering*, 2:1793–8201, Dec 2010.

[65] A. Srinivasan, R.D. King, S. H. Muggleton, and M.J.E. Sternberg. Carcinogenesis predictions using ilp, 1997.

[66] A. Srinivasan, R.D. King, S. H. Muggleton, and M.J.E. Sternberg. Structure-activity relationships derived by machine learning: The use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming, 1996.

[67] A. Srinivasan, R.D. King, S. H. Muggleton, and M.J.E. Sternberg. The predictive toxicology evaluation challenge, 1997.

[68] Nuno A. Fonseca, Max Pereira, Vítor Santos Costa, and Rui Camacho. Interactive discriminative mining of chemical fragments, 2010. Universidade do Porto.

[69] Vítor Santos Costa Nuno A. Fonseca and Rui Camacho. Logchem: Interactive discriminative mining of chemical structure. In *Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine*, BIBM '08, pages 421–426, Washington, DC, USA, 2008. IEEE Computer Society. doi:10.1109/BIBM.2008.45.

[70] Daniela Cardeano. Data mining em aplicações de desenho racional de fármacos. Technical report, Faculdade de Engenharia da Universidade do Porto, July 2014.

[71] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. Elsevier, 2006.

[72] Ruifeng Liu and Sung-Sau So. Development of quantitative structureproperty relationship models for early adme evaluation in drug discovery. 1. aqueous solubility. *Journal of Chemical Information and Computer Sciences*, 41(6):1633–1639, 2001. PMID: 11749590. URL: http://dx.doi.org/10.1021/ci010289j, arXiv:http://dx.doi.org/10.1021/ci010289j, doi:10.1021/ci010289j.